824/6.8.70.

824/6.8.70.

Vol. III Part I March 1950

478z 52

THE
BRITISH JOURNAL
OF
PSYCHOLOGY

Statistical Section

EDITED BY
CYRIL BURT AND GODFREY THOMSON

WITH THE ASSISTANCE OF

A. C. AITKEN
M. S. BARTLETT
W. G. EMMETT
M. G. KENDALL
D. N. LAWLEY
E. S. PEARSON
L. S. PENROSE
J. FRASER ROBERTS
A. RODGER
B. BABINGTON SMITH
W. STEPHENSON
P. E. VERNON

Managing Editor : H. G. MAULE

BUREAU OF EDNL. & PSYL. RESEARCH
LIBRARY

UNIVERSITY OF LONDON PRESS LTD.
WARWICK SQUARE, LONDON, E.C.4.

Price Fifteen Shillings

CONTENTS

		Page
VERNON, P. E.	An Application of Factorial Analysis to the Study of Test Items	1
RICHARDSON, L. F.	A Method for Computing Principal Axes	16
SEN, AMYA	A Statistical Study of the Rorschach Test	21
BURT, CYRIL	Group Factor Analysis	40
LAWLEY, D. N.	Factor Analysis by Maximum Likelihood : A Correction	76

The British Journal of Psychology (*Statistical Section*) is issued by the British Psychological Society. Three numbers will be published annually.

The subscription price is 30*s.* a volume of about 200 pages. Subscriptions should be sent to any Bookseller, or to THE MANAGER, University of London Press, Ltd., Warwick Square, London, E.C.4.

Papers for publication should be sent either to Professor Sir CYRIL BURT, Psychological Department, University College, London, W.C.1, or to Professor Sir GODFREY THOMSON, Moray House, Edinburgh, 8. Contributors will receive 20 copies of their papers free : extra copies may be had at cost price, if the order is given when the proofs are returned.

Authors are asked to send their papers in a form which is ready for submission to the printer : that is to say, the arrangement of the manuscript (including a brief list of contents at the head and a summary of results and conclusions at the end) should follow the general arrangement observed in the publications of the British Psychological Society.

Books for review should be sent to one of the Editors ; advertisements for insertion to the Manager, University of London Press, Ltd.

The British Psychological Society also issues quarterly *The British Journal of Psychology, General Section*, and *The British Journal of Medical Psychology*. Subscriptions, 40*s.* per annum for either journal, should be sent to the Cambridge University Press, Bentley House, Euston Road, London, N.W.1. The Society also publishes, jointly with the Training College Association, *The British Journal of Educational Psychology*. Subscriptions, £1 per annum, for this Journal should be sent to Methuen & Co., Ltd., 36 Essex Street, London, W.C.2.

Papers for publication in the *General Section* should be sent to Professor D. W. HARDING, the Psychological Laboratory, Bedford College, London, N.W.1 ; those for publication in *The British Journal of Medical Psychology* should be sent to Dr. J. D. SUTHERLAND, 2 Beaumont Street, London, W.1 ; those for publication in *The British Journal of Educational Psychology* to Emeritus Professor C. W. VALENTINE, The White House, Wythall, Nr. Birmingham.

Bureau Edni. [illegible] Research
6 8 50 COLLEGE

Vol. III Part II June 1950

THE BRITISH JOURNAL OF PSYCHOLOGY

Statistical Section

EDITED BY
CYRIL BURT AND GODFREY THOMSON

WITH THE ASSISTANCE OF

A. C. AITKEN
M. S. BARTLETT
W. G. EMMETT
M. G. KENDALL
D. N. LAWLEY
E. S. PEARSON
L. S. PENROSE
J. FRASER ROBERTS
A. RODGER
B. BABINGTON SMITH
W. STEPHENSON
P. E. VERNON

Managing Editor : H. G. MAULE

UNIVERSITY OF LONDON PRESS LTD.

WARWICK SQUARE, LONDON, E.C.4.

Price Fifteen Shillings

CONTENTS

		Page
Bartlett, M. S.	Tests of Significance in Factor Analysis . . .	77
Lawley, D. N.	A Method of Standardizing Group-Tests . . .	86
Lubin, A.	Linear and Non-Linear Discriminating Functions . .	90
Burt, C.	The Influence of Differential Weighting . . .	105
Thomson, G.	Note on Sir Cyril Burt's Paper on Differential Weighting	126
Burt, C.	A Reply to Sir Godfrey Thomson's Note . . .	127

The British Journal of Psychology (*Statistical Section*) is issued by the British Psychological Society. Three numbers will be published annually.

The subscription price is 30*s*. a volume of about 200 pages. Subscriptions should be sent to any Bookseller, or to The Manager, University of London Press, Ltd., Warwick Square, London, E.C.4.

Papers for publication should be sent either to Professor Sir Cyril Burt, Psychological Department, University College, London, W.C.1, or to Professor Sir Godfrey Thomson, Moray House, Edinburgh, 8. Contributors will receive 20 copies of their papers free : extra copies may be had at cost price, if the order is given when the proofs are returned.

Authors are asked to send their papers in a form which is ready for submission to the printer : that is to say, the arrangement of the manuscript (including a brief list of contents at the head and a summary of results and conclusions at the end) should follow the general arrangement observed in the publications of the British Psychological Society.

Books for review should be sent to one of the Editors ; advertisements for insertion to the Manager, University of London Press, Ltd.

The British Psychological Society also issues quarterly *The British Journal of Psychology, General Section*, and *The British Journal of Medical Psychology*. Subscriptions, 40*s*. per annum for either journal, should be sent to the Cambridge University Press, Bentley House, Euston Road, London, N.W.1. The Society also publishes, jointly with the Training College Association, *The British Journal of Educational Psychology*. Subscriptions, £1 per annum, for this Journal should be sent to Methuen & Co., Ltd., 36 Essex Street, London, W.C.2.

Papers for publication in the *General Section* should be sent to Professor D. W. Harding, the Psychological Laboratory, Bedford College, London, N.W.1 ; those for publication in *The British Journal of Medical Psychology* should be sent to Dr. J. D. Sutherland, 2 Beaumont Street, London, W.1 ; those for publication in *The British Journal of Educational Psychology* to Emeritus Professor C. W. Valentine, The White House, Wythall, Nr. Birmingham.

Vol. III Part III November 1950

THE BRITISH JOURNAL OF PSYCHOLOGY

Statistical Section

EDITED BY
CYRIL BURT AND GODFREY THOMSON

WITH THE ASSISTANCE OF
CHARLOTTE BANKS AND ARTHUR SUMMERFIELD
AND THE FOLLOWING EDITORIAL BOARD

A. C. AITKEN
M. S. BARTLETT
W. G. EMMETT
M. G. KENDALL
D. N. LAWLEY
E. S. PEARSON
L. S. PENROSE
J. FRASER ROBERTS
A. RODGER
B. BABINGTON SMITH
W. STEPHENSON
P. E. VERNON

Managing Editor : H. G. MAULE

UNIVERSITY OF LONDON PRESS LTD.
WARWICK SQUARE, LONDON, E.C.4.

Price Fifteen Shillings

CONTENTS

		Page
McLeish, J.	The Validation of Seashore's Measures of Musical Talent by Factorial Methods	129
Reid, D. D.	Night Visual Capacity and its Relation to Survival in Operational Flying	141
Reyburn, H. A., and Raath, M. J.	Primary Factors of Personality	150
Henrysson, S.	The Significance of Factor Loadings	159
Burt, Cyril	The Factorial Analysis of Qualitative Data . .	166
Lubin, A.	A Note on Sheldon's Table of Correlations between Temperamental Traits	186
Book Reviews	. . . .	190

The British Journal of Psychology (*Statistical Section*) is issued by the British Psychological Society. Three numbers will be published annually.

The subscription price is 30*s.* a volume of about 200 pages. Subscriptions should be sent to any Bookseller, or to The Manager, University of London Press, Ltd., Warwick Square, London, E.C.4.

Papers for publication should be sent either to Professor Sir Cyril Burt, Psychological Department, University College, London, W.C.1, or to Professor Sir Godfrey Thomson, Moray House, Edinburgh, 8. Contributors will receive 25 copies of their papers free : extra copies may be had at cost price, if the order is given when the proofs are returned.

Authors are asked to send their papers in a form which is ready for submission to the printer : that is to say, the arrangement of the manuscript (including a brief list of contents at the head and a summary of results and conclusions at the end) should follow the general arrangement observed in the publications of the British Psychological Society.

Books for review should be sent to one of the Editors ; advertisements for insertion to the Manager, University of London Press, Ltd.

The British Psychological Society also issues quarterly *The British Journal of Psychology, General Section*, and *The British Journal of Medical Psychology*. Subscriptions, 40*s.* per annum for either journal, should be sent to the Cambridge University Press, Bentley House, Euston Road, London, N.W.1. The Society also publishes, jointly with the Training College Association, *The British Journal of Educational Psychology*. Subscriptions, £1 per annum, for this Journal should be sent to Methuen & Co., Ltd., 36 Essex Street, London, W.C.2.

Papers for publication in the *General Section* should be sent to Professor D. W. Harding, the Psychological Laboratory, Bedford College, London, N.W.1 ; those for publication in *The British Journal of Medical Psychology* should be sent to Dr. J. D. Sutherland, 2 Beaumont Street, London, W.1 ; those for publication in *The British Journal of Educational Psychology* to Emeritus Professor C. W. Valentine, The White House, Wythall, Nr. Birmingham.

VOLUME III MARCH, 1950 PART I

478252

AN APPLICATION OF FACTORIAL ANALYSIS TO THE STUDY OF TEST ITEMS

By P. E. VERNON

Institute of Education, University of London

I. *Introduction.* II. *Data and Methods.* III. *Analysis of Items.* IV. *Analysis of Tests.* V. *Functional Content of Item Factors.* VI. *Functional Content of Separate Items.* VII. *Final Analysis of Formal Factors.* VIII. *Conclusions Regarding Factor Analysis.* IX. *Summary.*

I. INTRODUCTION

In an earlier article (Vernon, 25), it was suggested that factor analysis could be useful in analysing the homogeneity of content of test items, though the amount of work it involves limits its application to short tests. The present study was designed to discover how valuable would be the information obtainable about the items of a test of moderate length, namely, Progressive Matrices. This test was chosen, not with any intent to criticize, but because :

(*a*) it is so widely used that fuller understanding of what it measures is desirable ;

(*b*) answer sheets for the 20-minute version from 640 recruits, together with scores on 21 other tests, happened to be available ;

(*c*) the content of the test appears, from inspection, to be, if anything, more homogeneous than that of many other popular intelligence, attainment, and aptitude tests. Hence, if appreciable heterogeneity is found, we might need to reconsider the soundness of current methods of test construction.

Several researches have in fact indicated that the items of which certain tests are composed do measure diverse abilities. Burt (1), Wright (29), Burt and John (3), and McNemar (12) have shown that the Stanford-Binet or Terman-Merrill scales embody several group factors over and above their general content. In McNemar's 14 analyses at different age levels the mean first factor variance was close to 40 per cent. and the mean variance of his bipolars 13 per cent. But the latter were exaggerated by unreliable correlations from small groups, and amounted only to 10·5 per cent. in analyses where 200 cases were available.

The relative sizes of general and bipolar variances naturally depend on the heterogeneity of the subjects tested. But one would suggest that, in a representative group, the former should usually exceed the latter by at least four times. That this is a moderate demand is shown by Eysenck and Crown's (5) analysis of a 24-item attitude scale. There the general and bipolar factor variances were 48 per cent. and 5 per cent. Another research on the Stanford-Binet test, by Wright, yielded variances of 60·8 per cent. and 17·3 per cent. among ten-year olds. But as the last two of Wright's six bipolars were almost certainly non-significant, it would be fairer to take account only of the 12·9 per cent. variance of the four chief ones. The same criterion may be applied to the sub-tests of a battery. For example, the Army Summed S.G. is based on scores from five tests. When analysed in a representative group of recruits the first factor variance of these tests is 66·6, and three bipolars contribute 10·4 per cent. Had the ratio only reached, say, 3 to 1, we should be much more doubtful about employing a summed score.

An example of heterogeneity is provided by Guilford's (8) analysis of the ten sets of items in the Seashore Pitch Discrimination test, given to 300 students. He concludes that the more difficult sets (5 to 0·5 cycles) depend on entirely different factors from the easier sets (30 to 8 cycles). In spite of the test's apparently homogeneous material, it does not measure one and the same pitch ability throughout. An alternative possibility is that more able subjects use very different discriminatory methods from the less able. However Guilford notes that the two most difficult sets (1 and 0·5 cycles) are highly unreliable; and if these are omitted the first factor variance of the remainder is 57·8 per cent., bipolar variance 15·8 per cent. Indeed, with a less selected group (for Pitch Discrimination correlates appreciably with intelligence), the general factor in these sets might easily have satisfied our suggested criterion.

There is other evidence of 'difficulty factors.' Burt and John, in their analysis of Stanford-Binet X- and XII-yr. items, obtained a factor separating almost all the former from the latter items. Hertzman (10) studied correlations between the easier and more difficult halves of a battery of group-tests, and concluded that they measure somewhat different abilities. Ferguson (6) points out that spurious factors corresponding to the difficulty of different sets of items will be introduced if Pearsonian product-moment or point inter-correlations are analysed. Wherry and Gaylord (28) recommend substituting tetrachoric correlations, which are unaffected by item-difficulty; but in point of fact this had already been done both by Burt and Guilford in the researches just mentioned.

Another possibility is that the form of the items (open-type, analogies, classification, etc.) introduces unwanted factors. Guilford (7) claims that this is unimportant, and cites Smith's (18) investigation, where 14 verbal, numerical, and spatial tests in analogies, classification, and open-response form were compared. Although Smith's application of tetrad analysis failed to reveal any clear-cut formal factors, a re-analysis of his figures by the present writer, using group-factor technique, suggests the presence of a group factor with some 7 per cent. variance among multiple-choice tests (analogies *and* classifications) not present in open-response tests. The total variance of the factors based on the content of the tests among college students approximated 50 per cent. Most educational psychologists seem to regard an individual Stanford or Terman-Merrill I.Q. as having greater practical utility than an I.Q. from an equally reliable group-test, and this may be because the latter is distorted by some such formal factor, i.e., by the ability to do multiple-choice tests, which has been described by the writer elsewhere as 'test sophistication' (Vernon, 23).

Thurstone (22) and Shaefer (15) have suggested that the perceptual speed factor, *P*, appears only among tests composed of easy items. Though it was present in tests of matching or identifying printed letters and numbers, diagrams and pictures among Thurstone's (21) original college students, it failed to emerge consistently when similar tests were applied to schoolchildren, to whom such tests offered greater difficulty. It seems possible, again, that the word fluency factors, *W* or *f*, are differentiated from the verbal factor, *V*, largely by the ease of the verbal items involved, since several factorial investigations have failed to establish a clear separation between them. This is bound up with the whole question of the existence and nature of a speed factor in intelligence and other mental tests. For if such a factor is prominent it will presumably affect the performance of less able testees on an ordinary time-limit test more than it affects the more able; and it may affect the later items of such a test more than the earlier ones. In spite of the rejection of such a factor by Spearman and others, the investigations of Slater (16), Davidson and Carroll (4) and Tate (20) have proved beyond doubt that power and speed of work can be differentiated under appropriate conditions.

It might be thought that internal analyses which reveal heterogeneity among test-items can be disregarded, provided that scores on the test as a whole give useful predictions of some external criterion of educational or vocational success. But it is precisely because of the evidence accumulated during the war of irregular predictions at different test-score levels that the present writer has become concerned about heterogeneity.

One study of 600 naval radar operators yielded product-moment correlations of ·26 and ·35 for Progressive Matrices and Arithmetic tests with training marks. But when the 4 per cent. of failures and 96 per cent. of passes were contrasted, the tetrachoric coefficients were +·30 and −·13 respectively, showing that regressions were non-linear. The abilities measured by the Arithmetic test, or by the criterion marks, must have altered at the bottom end of the range. Similarly the Morse Aptitude Test, based on discrimination of Morse patterns, gave moderate correlations with Morse training results among telegraphists when product-moment correlations were used, but was extremely poor at picking out failures. In contrast the Morse Learning test (described by Vernon, 24) worked better at the lower than at the upper ranges. There was seldom time to investigate the significance of non-linearity, but it was definitely proven in one comparison, by Dr. Fraser Roberts, of the Matrices test with the naval T2 battery.

Clearly there is room for much more detailed analysis of what our tests measure, for determining the extent to which all items depend on the same factor or combination of factors, and whether heterogeneity of content or form is responsible for variations in predictive value.

The Progressive Matrices test[1] was originally devised as a measure of *g*, but it usually obtains small spatial or *k*-loadings, among adults, as shown by the writer elsewhere (24, 26). Its five sets of 12 items employ similar material, but differ slightly in their approach. Raven (13) describes them as :

(*a*) Continuous patterns,
(*b*) Analogies between pairs of figures,
(*c*) Progressive alterations of figures,
(*d*) Permutations of figures,
(*e*) Resolution of figures into constituent parts.

Each set starts with very easy items, so that separate instructions are not required, and each finishes with moderately, or very, difficult items. Only one study—to the writer's knowledge—has been made of the separate sets. Rimoldi (14) factorized these and 14 other tests, which had been given to 138 children of mixed sex, and age-range 8 to 15 years. Seven factors were extracted and rotated, and as might be expected the results are merely confusing. However, the unrotated factors indicate common variance in the Matrices sets of some 55 per cent., together with bipolars, which differ in the sets, of some 10 per cent.

Spearman (19) and others have suggested that different subjects might answer the items by different methods, for example, synthetically or analytically. It seems likely also that some would resort to spatial imagery, and others might employ logical verbal analysis. Heterogeneity of method in the persons would, on Burt's reciprocity principle, presumably imply heterogeneity of item-content in the tests. It was decided in the present study both to analyse the items (not the sets) among themselves, and to compare each item with other tests which would provide measures of well-established factors—*g*, *v*, *n*, *k*, mechanical information and perceptual speed.

II. DATA AND METHODS

The following battery of tests was assembled by Science 4, the Air Ministry, and given experimentally in the summer of 1948 to 640 National Service male recruits, all aged close to 18 years. Grateful acknowledgements are due for permission to use the results. Fuller descriptions of the tests may be found in Vernon and Parry (27).

---

[1] The original 1938 version was used in this investigation. It has, however, since been claimed that the later version, 1947, Set II, already meets many of the weaknesses described below. The items are arranged in groups of four, instead of twelve, according (*a*) to argument and (*b*) to order of difficulty. It would be preferable for this test to be used in future large-scale investigations of adult intelligence.

Arithmetic A. Short problems.
Calcs. A and Calcs. C. Four rules and mechanical arithmetic.
V-4. Stephenson's mixed vocabulary, English, and reading comprehension test.
V-V. A new vocabulary and comprehension test.
Gen A-1. Crossing out wrongly spelled words.
Gen A-2. Comprehension of paragraphs.
SP 14. A.T.S. Spelling ; identifying the correct spellings from several misspelled versions.
SP 21. Army Instructions or Clerical test.
SP 119. Giving the readings of pointers on scales and dials, and the co-ordinates of points on rectangular and polar graphs.
Ins-A. Reading aeroplane panel dials.
Obs-C. Matching aircraft silhouettes.
Progressive Matrices. Raven's 1938 test, with 20-minute time limit.
G-5. Stephenson's Matrices test, modified, with multiple-choice responses.
K-6. Stephenson's spatial test, including paper formboard, copying figures onto dotted space, and drawing mirror images.
SP 4. National Institute of Industrial Psychology's spatial judgment test, Squares.
Group Test 80. N.I.I.P. test of recognition of shapes turned round or over.
SP 122. Army revision of Bennett Mechanical Comprehension.
Mec-B. Diagrams of machines, comprehension questions.
Mec-C. Multiple-choice mechanical information questions.
SP 117 E and M. Questions on everyday electrical and mechanical matters, open-response.

Table I gives the 20-minute Matrix norms believed to be correct for 18-year males in the general population, and the distribution in our sample.

TABLE I. PROGRESSIVE MATRICES DISTRIBUTIONS

Score	18-yr. Population Per cent.	R.A.F. Sample Per cent.
50+	10	17·8
46+	20	25·7
42+	20	25·6
36+	20	21·6
28+	20	8·7
27 and under	10	·6

Unfortunately the sample is considerably above average ; and, though it is as heterogeneous as those normally encountered in educational or vocational investigations outside the Army, the Standard Deviation of scores is only 6·18. In more representative Army groups the Standard Deviation is about 8·50. Not only does this reduce the common-factor variance, but it also means that few recruits fail the easier items. Actually in 30 items out of 60 the number failing amounted to less than 10 per cent.

Tetrachoric correlations can still be obtained between items with very low (or very high) failure rates, but they are extremely unreliable. Thus the S.E. of zero r, according to Kelley's formula (11, No. 211) is ·062 when both items are passed by about 50 per cent. of a group of 640. But for the correlation between items C 1 ($p = \cdot970$) and E 12 ($p = \cdot041$), it rises to ·223. This difficulty will always arise in a test covering a wide range of ability. It might be met by taking an abnormally heterogeneous population, or, following McNemar, by making several overlapping analyses among low-grade and high-grade populations. However, most tests are not, like Matrices, intended to cover Mental Ages from 8 up to superior adult, and their items do normally lie between 10 per cent. and 90 per cent. difficulty.

The 18 easiest items were omitted altogether ; and the p values (difficulties) of the remaining 42 appear in Table II, Col. 2. The S.E. of zero r between each item and an item whose $p = \cdot50$ is given in Col. 3. For any other pair of items, the S.E. of zero r is given by the product of these two entries, multiplied by 16·1. (For example, C 1 with E 12 : $\cdot123 \times \cdot112 \times 16{\cdot}1 = \cdot223$.) The relative difficulties parallel very closely those quoted by Slater (17) for a group of 571 Army recruits, the rank order correlation between them being ·946.

Scores on the 22 tests were split at the median. It might be preferable to correlate each Matrices item with tests split at the corresponding difficulty level, but this would have enormously increased the labour, since the tests themselves would then have had to be inter-correlated and factorized at every level, or at least at several different difficulty levels. It appeared better to maintain a single level for the tests in order to provide a stable frame of reference factors with which not only the medium, but also the very easy and very difficult, items could be compared.

Tabulation of the 2,016 correlations between 42 items and 22 tests was done by hand, since punching was not readily available. Tetrachorics were read off from graphs. In view of the large S.E.s, and in order to save time in what was intended solely as an exploratory investigation, most of the subsequent calculations were done to two decimal places only.

III. ANALYSIS OF ITEMS

It is usually unwise to factorize two 'universes' of variables, such as test items and test scores, in a single matrix. Separate analyses were therefore made of the inter-item correlations, the inter-test and the item-test correlations. The first was carried out by the Centroid method, and the four factors extracted are shown in Cols. 4–8 of Table II. Communalities were guessed after totalling the columns at each stage; neither successive approximation nor the highest r in each column was used.

It is difficult to evaluate significance, but if we may assume (with McNemar) that a factor is equivalent to a variable split at the median, then the figures in the third column of Table II give us the S.E. of zero r for first-factor loadings. For the sth factor, the S.E., according to Burt and Banks (2) is $\sqrt{\frac{n}{n-s+1}}$ times larger. The writer's practice is to regard a factor as significant if half the loadings exceed 2×S.E. Here 19 out of 42 fourth-factor loadings reach this level. On Guilford and Lacey's (9) criterion also, the fourth factor is close to the borderline of significance. Hence it did not appear justifiable to extract a fifth or further factors. Nevertheless the total communality obtained, 39·7 per cent., is a conservative estimate; and it is likely that more efficient techniques of analysis, such as Maximum Likelihood, might raise it to perhaps 42 per cent.

This analysis is, by itself, not very informative. Most of the item inter-correlations can be accounted for by Factor I, with 26·0 per cent. variance, but the bipolars cover another 13·7 per cent. If rotation was attempted to factors with positive or near-zero loadings, the general factor would largely disappear, and independent factors corresponding quite closely to Raven's sets of items ($A+B$, C, D, and E) would be obtained. Most American factorists would almost certainly conclude that each set is measuring a different ability. However the heterogeneity is not really so serious, for two reasons, the first being the attenuation of Factor I by the homogeneity of the population. The total variance of 39·7 per cent. can be corrected by Kelley's formula (11, No. 178) to the value expected in a population with Standard Deviation 8·50. It then becomes 68·2 per cent., and as most of the increase would occur in Factor I, the variance of this factor might reach 54·5, that is, more than double the obtained value. Note that this figure is just about four times 13·7. Hence the Matrices test appears to satisfy the criterion of homogeneity suggested above (unless the bipolars also swelled in a representative group). Nevertheless it is instructive to realize how small the general factor may be, and how relatively great the heterogeneity, in a test applied to a moderately selected population.

Secondly, it is probable that factors with all-positive loadings would not give a true picture of item structure, and that at least one factor should be bipolar. Consideration of the answer sheets and item-correlations strongly suggests that many subjects spend so much time on the more difficult early items that they scarcely

TABLE II. RESULTS OF FACTOR ANALYSES OF MATRICES ITEMS

Column 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
			Centroid Item Factors					Functional Factor Content							Formal Factor Content					
																	Group			
Item	p	S.E. at $p = ·50$	I	II	III	IV	h^2	g	n	v	Technical	k	Perceptual	h^2	Bi-polar	General	B	C	DE	h^2
A 11	·753	·067	·28	−·08	−·17	−·19	·15	·16			·14	·30		·14	+·13	·03	·12			·03
12	·920	·090	·30	−·06	+·20	−·03	·14	·34						·12	+·12	−·10			·24	·08
B 6	·947	·102	·33	−·03	−·34	−·35	·35	·23			·21	·32		·20	+·14	−·06	·23			·07
7	·872	·079	·37	−·24	−·23	−·14	·27	·23	·17			·14	·23	·16	+·29	·15	·20			·14
8	·833	·095	·30	−·34	+·28	−·14	·30	·22			·37			·18	+·16	·15	·03			·05
9	·706	·071	·28	−·25	−·12	−·13	·17	·22						·05	+·27	·11	·37			·21
10	·925	·092	·49	−·32	+·06	−·47	·57	·30		·31		·30	·25	·34	+·30	·14	·51			·36
11	·795	·070	·40	−·19	−·40	−·29	·44	·23	·18		·13			·10	+·36	·27	·41			·36
12	·601	·063	·48	−·15	−·14	−·18	·31	·38				·29		·23	+·23	·11	·25			·12
C 1	·970	·123	·48	+·10	−·30	+·10	·34	·25				·41	·22	·28	+·10	·23		·36		·19
4	·912	·088	·50	+·17	−·26	−·06	·35	·37				·27	·20	·25	−·02	·09		·48		·25
5	·941	·099	·57	+·03	−·23	−·19	·42	·44			·26	·34		·38	+·05	·12		·20		·05
6	·842	·075	·30	+·25	−·22	+·07	·21	·37	·17					·17	−·15	−·09		·42		·21
7	·937	·098	·60	+·06	−·08	−·11	·38	·40				·34		·28	+·09	·38		·15		·17
8	·689	·065	·44	+·11	−·16	+·13	·25	·25				·14		·08	+·05	·30		·36		·22
9	·742	·067	·38	−·05	−·17	+·16	·20	·33						·11	+·04	·19		·22		·09
10	·471	·062	·38	+·15	−·21	+·15	·23	·33				·24		·17	−·10	·05		·03		·02
11	·331	·064	·30	+·13	−·27	+·23	·23	·19	·12	·15				·07	−·01	·14		·36		·15
12	·061	·098	·53	+·22	−·15	+·18	·38	·67				·28		·53	−·19	·00				·04

D 2	·927	·092	·59	−·22	+·22	+·34	·56	·30		·11	·12	·14	·45	·34	+·17	·20	·66	·51
3	·923	·092	·55	−·41	+·14	+·29	·58	·36		·13		·31	·32	·35	+·28	·00	·69	·56
4	·902	·085	·51	−·43	−·06	+·26	·52	·38			·15			·17	+·34	·49	·16	·38
5	·948	·104	·64	−·28	+·34	+·19	·64	·44	·17		·12			·24	+·20	·17	·88	·85
6	·897	·084	·69	−·36	+·09	+·07	·62	·56			·20			·35	+·27	·51	·10	·34
7	·800	·071	·57	−·36	+·16	+·01	·48	·34			·12			·13	+·26	·54	·22	·41
8	·802	·071	·61	−·18	+·14	+·07	·43	·35		·13		·24		·20	+·15	·28	·52	·37
9	·637	·063	·54	−·22	+·04	−·12	·36	·40			·14	·24		·20	+·13	·27	·14	·11
10	·579	·062	·67	−·02	−·03	+·21	·49	·50			·22	·21		·34	+·07	·40	·08	·17
11	·205	·070	·03	−·13	−·01	+·16	·04	·00					·15	·02	+·00	−·09	·30	·10
12	·084	·089	·42	−·28	+·06	+·20	·30	·32		·27				·18	+·17	·34	·07	·15
E 1	·786	·070	·57	+·05	+·35	−·16	·48	·46			·14	·23		·28	−·20	·21	·15	·10
2	·626	·063	·72	+·29	+·27	+·14	·70	·68			·13			·48	−·34	·40	·07	·27
3	·643	·064	·64	+·18	+·08	−·18	·48	·61			·17	·12		·42	−·23	·42		·23
4	·439	·062	·75	+·25	+·11	+·07	·64	·70			·14	·11		·52	−·26	·34	·06	·19
5	·444	·062	·77	+·28	+·08	−·01	·68	·70						·49	−·21	·81		·70
6	·369	·063	·62	+·31	+·08	+·05	·49	·59			·12	·15		·39	−·31	·40		·26
7	·111	·082	·44	+·53	+·19	−·06	·52	·41		·11		·35		·30	−·45	—		·20
8	·056	·101	·02	+·31	+·31	−·14	·21	·10		·27		·29	·38	·31	−·55	—		·30
9	·282	·066	·48	+·35	−·10	−·18	·40	·45		·13		·36		·35	−·29	—		·08
10	·242	·068	·70	+·32	+·04	−·09	·60	·70			·18	·19		·56	−·27	—		·07
11	·204	·070	·63	+·25	+·14	−·16	·51	·63		·10		·30		·50	−·26	—		·07
12	·041	·112	·27	+·28	+·35	+·03	·28	·60		·30				·45	−·52	—		·27
Variance per cent. ...			26·0	6·3	4·1	3·3	39·7	18·4	·3	1·0	1·4	4·3	1·6	27·0	5·9	7·5	9·2	22·6
Functional Content of Item-Factors I								·86		·13	·10	·33		·88				
II								·16	·07	·11		·22	·11	·10				
III										·26	·08	−·15	·17	·13				
IV									−·15		·09	·28		·11				
Variance per cent. ...								19·4	·1	·8	·3	3·5	·2	24·4				

have time to tackle Sets *D* and *E*, whereas others guess freely and have more time to score on the later sets. If so, some groups of items really are negatively correlated with others. Doubtless this phenomenon arises largely because the test is applied with a time limit, which was not intended by its author. The moral is clear: tests given with insufficient time should have the items arranged in order of difficulty throughout, and if the items fall into distinctive sets, they should be arranged in omnibus or cyclic form. Otherwise the total scores of some subjects will be based to too great an extent on some of the sets, and of other subjects on other sets.

There is unfortunately no direct means of telling which bipolar originates in this manner. Order of difficulty gives some clue, although the subjective difficulty felt by the average testee does not necessarily correspond closely to the proportions of failures, $1-p$. By rotating Factors II and III through $\tan^{-1} \cdot 5$, a bipolar is obtained whose loadings give a tetrachoric correlation of ·73 with p. This is the best approximation we can get to a difficulty factor. It is shown in Col. 16 of Table II.

We cannot decide, either, whether our factors derive from the form or the functional content of the items. Probably all of them are mixtures. As already pointed out, the bipolars differentiate Raven's sets; but it may well be that these sets differ in function (e.g., dependence on k) as well as in form. Although Factor IV yields mixed positive and negative loadings for at least two sets, *C* and *E*, we cannot assume that this represents differentiation by function. It may merely separate the more from the less difficult items within these sets. Thus the six most difficult items in Set *C* all have positive loadings and three of the four easier ones have negative. Similarly Items *E* 7–11 are all negative.

How far do general and bipolar variances differ with difficulty? Table III gives the results for the easiest, medium, and most difficult items.

TABLE III. FACTOR VARIANCES OF EASY AND DIFFICULT ITEMS

Factor	I	II	III	IV	h^2
22 easy items, $p = \cdot 7$ to $1 \cdot 0$ ...	23·6	5·8	5·3	4·3	39·0
11 medium items, $p = \cdot 3$ to $\cdot 6$ ...	35·2	4·3	2·5	2·2	44·2
9 difficult items, $p = 0$ to $\cdot 2$ ...	20·5	9·8	3·5	2·1	35·9
All items	26·0	6·3	4·1	3·3	39·7

Apparently the medium ones have the greatest homogeneity and least bipolar variance. The easy and difficult ones both have higher than average bipolar content, but their low Factor I content is even more noticeable. The tendency for extreme items to be weaker than medium ones (even when consistency or validity is determined by tetrachoric or biserial correlation) was pointed out in the writer's survey of item-analysis techniques. The figure for the most difficult Matrices items is especially low because they include two items that are almost useless for adults, *D* 11 and *E* 8.

Any information about test items, based on general factor loadings, could of course be obtained much more simply by the stock item-analysis techniques. On the whole, then, factor analysis of items by themselves hardly seems worth while, provided that the internal consistency or reliability of the test is sufficiently high to ensure the dominance of a general factor. The bipolars have told us nothing, except that it is probably a mistake to give Matrices with a time limit. Conceivably they would be more useful in the case of a test where their meaning could be more readily identified by studying the content of the items they differentiate. Yet even in McNemar's Terman-Merrill analyses, interpretation was often difficult. An important reason for this is that the bipolar loadings are inevitably low in reliability when based on tetrachoric item-correlations.

IV. ANALYSIS OF TESTS

The inter-test correlations were analysed by group-factor methods, a previous centroid analysis of the same battery (slightly abbreviated) having revealed the general structure of the tests (Vernon, 26). *G* loadings were chosen to reduce to near-zero the correlations between sets of tests unlikely to have any group factor in common, viz.: Calcs. A and C; G-5 and Matrices; Mec-C, 117 E and M. Similarly V-4, Gen A–1, and SP 14 should be independent (apart from *g*) both of G-5 and Matrices, and of K-6, SP 4, and Group Test 80. The *g*-loadings of other tests, such as Obs-C or Scale Reading, were obtained by summing their correlations with the block of tests to which they were least related, and dividing by the summed *g*-loadings of this block. Admittedly there is an element of subjective judgment in this procedure, but it was guided both by the centroid analysis and by many previous analyses of most of the tests in comparable populations.

TABLE IV. FACTOR ANALYSIS OF TESTS

Test	*g*	*n*:ed.	*v*	Technical	*k*	Perceptual	h^2
Arithmetic A	·78	·44	(·16)	(·09)			·836
Calcs. A	·78	·59					·956
Calcs. C	·77	·53					·874
V-4 Verbal	·75	(·13)	·47				·801
V-V Vocabulary and Comprehension...	·63		·58	(·13)			·750
Gen A-1 Spelling	·67	(·19)	·49				·725
Gen A-2 Paragraph Comprehension ...	·70		·38				·634
SP 14 Spelling	·58	(·32)	·69			(·15)	·937
SP 21 Instructions	·84		·28			·38	·928
SP 119 Scale and Graph Reading ...	·88	(·09)		(·19)		·24	·876
Ins-A Dial Reading	·90		(·13)			·26	·895
Obs-C Matching silhouettes	·53			(·19)	(·34)	·22	·481
Progressive Matrices	·84				(·27)		·779
G-5 R.A.F. Matrices	·90				(·28)		·888
K-6 R.A.F. Spatial	·74			(·13)	·64		·975
SP 4 Squares	·51			(·21)	·38		·448
Group Test 80	·56			(·23)	·53	·12	·662
SP 122 Practical Problems	·56		(·18)	·34	(·43)	(·28)	·725
Mec-B Mechanical Comprehension ...	·44		(·39)	·51	(·18)	(·27)	·711
Mec-C Mechanical Information ...	·62		(·28)	·50			·712
SP 117 E Electrical Information ...	·59		(·28)	·47			·647
SP 117 M Mechanical Information ...	·50		(·30)	·59			·688
Variance per cent.	48·9	4·5	9·0	6·4	6·0	2·3	77·1

After removing *g*, the prominent residuals between the three arithmetic tests gave the *n* (number) loadings shown in Table IV, by Spearman's general factor formula. Certain other tests had smaller positive residuals with these three, and their *n* saturations are shown in brackets. Clearly this is as much an elementary education as a pure arithmetic factor. Spelling obtains a considerable loading, but the verbal comprehension tests V-V and Gen A-2 are free from it. Five verbal tests similarly yielded the next, *v*, factor and this entered to some extent into the five mechanical tests. Although this finding is unusual, it is quite plausible, since the mechanical tests obviously involve the understanding of printed material.

The next factor probably represents mechanical information and technical education, since Arithmetic, Scale Reading, Obs-C and the *k*-tests all show small loadings. The three spatial tests have a prominent additional group factor, which is present in mechanical comprehension (SP 122 and Mec-B) though not in information, also in G-5 Matrices and Obs-C. Finally there is slight evidence to support Thurstone's factor of perceptual speed in clerical and other tests which involve rapid matching or identification of verbal, numerical, or pictorial material. Actually its make-up is more reasonable in this than in the previous, centroid analysis, and though its reliability is low, it is probably significant. By Burt and Banks' formula, with $n=5$ and $s=3$ (since *g* and 1 or 2 group factors have been extracted from its component tests), the S.E. of a zero loading is ·081. Four of the five loadings exceed twice this figure.

Table IV shows that *g* covers 48·9 per cent. of test variance and the five group factors together 28·2 per cent., total 77·1 per cent. Corresponding figures for the centroid analysis were 43·8 per cent., 27·4 per cent., and 71·1 per cent., though comparisons are hardly legitimate both because the centroid analysis combined together certain pairs of tests here analysed singly, and because it was based on product-moment coefficients. It may be noted that neither analysis gave any indication of formal factors, although the tests do include open-response and various types of multiple-choice items, and although some are predominantly ' power ' and others are ' speed ' tests. Probably the functional content quite outweighed any grouping according to form.

V. FUNCTIONAL CONTENT OF ITEM FACTORS

The overlapping of items and tests could be approached in a number of ways. Here, each test was first treated as an additional Matrices item, and its correlations found with the four centroid item factors. These correlations were then treated as those of an additional test in the group factor analysis. The results appear at the foot of Cols. 9–15 in Table II.

As might be expected, the first item factor is largely composed of *g*. Its loading of ·86 corresponds nicely with the loading of ·84 for Matrices total score in Table IV. But it embodies a considerable amount of *k* also, and shows loadings with *v* and the technical factor which are probably insignificant. Only 12 per cent. of this factor cannot be identified functionally.

The bipolar factors are more complex. Their group-factor loadings are mostly non-significant, though they range up to ·28. Had the factors been appropriately rotated, it is conceivable that one might have been fairly highly *k*-saturated, another might have embodied all the *v* or *n* content, and so on. But the total functional content is so small that it seemed hazardous to attempt this. The total content (in so far as it can be identified by our reference battery) is $\cdot 260 \times \cdot 88 + \cdot 063 \times \cdot 10 + \cdot 041 \times 13 + \cdot 033 \times \cdot 11 = 24{\cdot}4$ per cent. The difference between this and the communality of the item factors, 39·7, presumably represents formal factors. That is to say, the combined variance of the spurious factor introduced by the imposition of a time limit, and of common factors in all items or in groups of items which have no identifiable functional meaning, amounts to some 15 per cent. (It may be rather larger if, as suggested on p. 5, the 39·7 is an under-estimate.) A further attempt to discover, and assess the size of, the component elements of this 15 per cent. is described in Section VII.

VI. FUNCTIONAL CONTENT OF SEPARATE ITEMS

The tests were divided into five blocks corresponding to the five group factors (omitting G-5 and Matrices total), and the *g*-content of each item found from the block with which it gave the lowest correlations. Its group-factor content was then determined from its residuals. Table II, Cols. 9–15, lists only those group-factor loadings which are at least as large as their presumed S.E.s.

As would be expected, the *g*-loadings parallel closely the Item-factor I loadings, but are mostly rather smaller, since only 74 per cent. (i.e., $\cdot 86^2$) of Factor I is *g*. Loadings on *k* range from ·41 to zero, hence we can almost certainly conclude that some items do demand more spatial visualization than others. Sets *C* and *E* tend to be more spatial than *B* and *D* ; and this could have been deduced from the *k*-loadings of the item bipolars (thus Factor II, in which Sets *C* and *E* have positive loadings, correlates ·22 with *k*). It does not seem possible to tell, from inspection, which kinds of items are likely to be most spatial.

Several items, particularly in Set *E*, obtain *v*-loadings. Though unexpected, it is conceivable that such items demand a species of logical analysis in which the verbally minded have an advantage. Seven out of the 11 most difficult items in the test are among these *v*-loaded ones ; this corresponds to an r_t of ·74 between difficulty and *v*. The *n*-factor shows small loadings for five items only. This may be chance, but inspection suggests that they may involve counting or other mathematical thinking to a greater extent than the rest of the items. The perceptual loadings are interesting since they are mostly (though not invariably) confined to very easy items—thus confirming Thurstone and Shaefer's hypothesis. Four of the eight are among the eight easiest items, yielding an r_t of ·63. Obviously this association is statistically unreliable, but it is suggestive. Conceivably if we had been able to analyse more of the easy *A* and *B* items, we should have found this factor playing a more prominent part.

The most puzzling result is the number of items showing technical-factor loadings. This ought not to occur, since the test as a whole is free from the factor. Possibly it arises merely because mechanical comprehension tests are *k*-saturated, and it would have been avoided had the *k* group factor been extracted before instead of after the technical factor. By no stretch of the imagination can items *B* 8 and 6, C 5, *D* 10 and 6 be seen to involve technical knowledge any more than other adjacent items. Most of the remaining loadings are very small, as also are the technical loadings of the item bipolar factors.

When the variance of *g* and the group factors in all the items is totalled, it amounts to 27·0 per cent. That it should exceed 24·4 per cent. (the functional content of the item factors) is quite correct, since the specificity of an item might well consist, in part, of *g* or of some familiar group factor. This would explain, also, why some of the communalities in Col. 15 exceed those in Col. 8, though probably the main reason is, simply, unreliability.

Perhaps the chief finding of this section is that the functional content of Matrices items (as distinct from their formal factor loadings) does not alter markedly as one progresses. There will be a slight tendency for very low scores to be based on perceptual factor to a greater extent than average scores, and a rather more marked tendency for very high scores to be *v*-loaded. Also, of course, the average *g*-saturation tends to rise from beginning to end. But generally speaking scores at all levels covered by this investigation represent much the same mixture of *g* and *k*. From this point of view the Matrices test probably shows rather better homogeneity of content than does the Terman-Merrill scale. On the other hand Terman-Merrill is likely to be much more free from formal factors, for these undoubtedly do have different effects on different levels of Matrices scores.

VII. FINAL ANALYSIS OF FORMAL FACTORS

Since it seemed impossible to resolve the original item factors any further, a fresh attempt was made by first extracting the *g* and group factor content from the original item inter-correlations. The residuals, which were presumably now linked by formal factors only, were then subjected to centroid analysis and three factors extracted. Their variance was 8·9, 4·2, and 4·7, total 17·8 per cent. This is a little higher than the 15 per cent. previously allowed, but naturally so because the analysis has gone much further. Instead of four item factors, we have removed *g*, five group factors, and three more bipolars.

These factors are not shown here, since all attempts to rotate them into logical structure proved fruitless.[1] They did however suggest that the structure consists of:

(i) the bipolar difficulty factor;
(ii) a general factor in almost all items, representing the test's specificity;
(iii) separate group factors in the *B*, *C*, and *D* sets, and possibly two group factors in *E* 1-6 and *E* 7–12.

Here we have seven factors. It is hardly to be wondered at that the original four item factors, in which not only these but also the six content factors were amalgamated, were difficult to interpret.

That No. (ii) exists, i.e., that there is some positive overlap among all items over and above their common *g*, *k*, etc., content, was proved by studying the residuals of each item with Matrices total (after removing *g* and *k*). The mean variance of these residuals was 4·5 per cent.

Since neither centroid nor group-factor techniques can readily cope with a general + bipolar + group-factor structure, it was eventually decided to use the rotated Item factors II + III (cf. p. 6) as an approximation to the bipolar difficulty factor, and to remove this first. The residuals were then analysed by Burt's group-factor technique. *E* 7–12 no longer showed any common variance, and *E* 1–6 overlapped closely with Set *D*, hence three, instead of five, group factors were sufficient. Item *A* 11 fitted in with the *B* set, but *A* 12 more closely resembled the *DE* set.

The results appear in Cols. 17–21 of Table II. A few of the loadings are abnormal, for example, *E* 5's general factor saturation of ·81 and *D* 5's group factor of ·88. But at the tail end of such complex manipulations of rather unreliable inter-item correlations this is only to be expected. The general factor loadings for the 36 items agree quite closely with the Matrices total score residuals ($r_t = \cdot 76$). What is more important, the summed communalities for the six functional and five formal factors agree quite well with those of the four item factors ($r_t = \cdot 82$), though they tend to be over-large.

The variance of the difficulty bipolar is 5·9, of the general factor 7·5, and the group factors 9·2, total 22·6. If the two abnormal loadings, just mentioned, are omitted, the general and group factor loadings drop to 5·9 and 7·4, the total to 19·2. This is still too large. We know that the general factor variance should be near to 4·5, and that the total should not much exceed 15. But considering that we have isolated 11 factors in place of four, the exaggeration is not unduly great. Our main conclusions are that the bipolar + general + group factors give a fair approximation to the structure of the formal factors, and that these three components are roughly equal in variance in the average item. There may be other formal factors linking smaller sets, or pairs, of items, but they are probably of negligible size.

Broadly speaking then an average Matrices item is determined by: *g* 19 per cent., *k* or technical 5 per cent., *v* or *n* 1½ per cent., perceptual 1½ per cent., positive or negative difficulty factor 5 per cent., Matrices factor (common to all items but to no other tests) 5 per cent., Set group factor 5 per cent. The total is 42 per cent., item specificity 58 per cent. Had the test been given to a more representative group of adults, the last figure would drop to 32 per cent. and, so far as we can guess, most of the increase would have occurred in the *g*-variance. But as we do not know what selective variables are operating in our population, this is uncertain.

---

[1] The first was positive for all items except *E* 7–12 and *C* 12. It was presumably a mixture of (i) and (ii). But it also gave very high loadings to the *D* set of items, i.e., it represented the group factor (iii) *D*. This explains why it could not readily be resolved by rotation.

VIII. CONCLUSIONS REGARDING FACTOR ANALYSIS

This exploratory investigation has been crude and not very accurate, and it has been hampered by inadequate tests of significance. (For example it is quite impossible to judge whether Matrices items differ significantly in perceptual factor loadings.) But this would seem to be inevitable when tetrachoric coefficients of poor reliability are analysed, and when the number of variables is far too large for successive approximation to communalities, or for the application of more efficient iterative techniques. In extracting the centroid factors, doubts frequently arose as to which variables to reflect, and there was certainly more than one possible solution which satisfied the criterion of positive column totals (disregarding the communalities). Wrigley (30) has recently discussed this problem and concluded that the factorist should aim to maximize the variance of each bipolar by reflecting tests with small loadings in such a direction as to increase the column totals of tests with large loadings, even if this does not produce all-positive totals.

In spite of these drawbacks the study has thrown much interesting light on test construction problems in general, and has permitted for the first time an assessment of the relative variance of functional and formal factors in a widely used group test. Whether the information obtained is of sufficient value to justify the work involved when constructing a new test is more dubious. There one often tries out a hundred items to pick the 50 best, and factorization would be out of the question. The analysis of items by themselves tells us very little which could not have been learnt more simply from stock item-analysis techniques. It appears to be capable of revealing only major defects of structure like the difficulty factor in Progressive Matrices and the prominent variations of functional content in the Terman-Merrill scale.

The inclusion of a battery of reference tests in the analysis constitutes a powerful tool, though again one that is very time-consuming. At least three tests must be given for each group factor that may be involved. Moreover, very large numbers of testees are essential before we can be certain, say, that an item designed to measure *g* is overloaded with *k*. If the present writer had been commissioned to revise the Matrices test, he would be extremely chary of eliminating or altering any of the items on the basis of the results given in Table II. Probably he would regard as the weakest items those that showed the largest proportions of functional group-factor and formal-factor variance, in other words the smallest *g*-variance. But this *g*-variance could have been obtained very easily by ordinary item analysis against a short battery of tests providing a reliable external criterion of *g*. So that here again factor analysis seems superfluous.

One way of making such research more manageable would be to combine items into a limited number of groups, each group being as homogeneous as possible in form, difficulty, and content. For example, the four easiest, four medium, and four most difficult items in each of Raven's sets might be treated as single variables. It would still be necessary to use tetrachorics. But the analysis of 15 item-groups would be a very different matter from analysing 60 items. It would be of great interest to do this with some of the popular group tests which contain several distinct sub-tests, or which follow the cycle-omnibus pattern, for example, Otis, Simplex, Cattell, Moray House, or Duplex.

IX. SUMMARY

1. The application of factorial analysis to a set of test items (or to the sub-tests of a battery) will help to show, first, whether the content of the test is homogeneous, or whether different items or sub-tests measure largely distinct abilities. It is suggested that the first or general factor running through all the items or sub-tests should obtain at least four times the variance of all the significant bipolar factors (before rotation), when analysed in a representative sample, if the test is to be regarded as adequately homogeneous.

2. Secondly, it may indicate whether unwanted 'formal' factors are present in some or all of the items. These include factors corresponding to the form in which the item is set, speed factors, or a factor contrasting difficult with easy items, etc. Such factors probably play an unduly large part in most multiple-choice tests. They are contrasted with 'functional' factors, which have useful psychological meaning, such as *g* and the recognized group factors, *v*, *n*, *k*, etc. It is possible that some of the familiar factors such as *W* and *F* (fluency) and *P* (perceptual speed) are primarily formal.

3. Such analyses are extremely time-consuming when the number of items is at all large, and the conclusion is reached that they contribute little which could not have been learnt from stock item-analysis techniques, unless it be certain grosser defects of the test's construction. Items must be inter-correlated by tetrachoric technique, and the resulting coefficients and factors tend to be very unreliable. It is suggested that analyses of groups of items, homogeneous in form, difficulty, and content, would be more economical. Several overlapping analyses at different difficulty levels may be essential for a test covering a wide range of ability.

4. Functional and formal factors cannot be distinguished unless the items are analysed along with an extensive battery of reference tests which will identify the item *g* and group-factor content. This too is seldom practicable, and is hardly necessary if ordinary item analysis is carried out against an external criterion.

5. Nevertheless heterogeneity and formal factors constitute a serious problem, as shown by the non-linear regressions frequently noted in following up Army and Navy tests during the war. An exploratory study was therefore made of one widely used test, the Progressive Matrices (1938), given with 20-minute time limit.[1]

6. Forty-two of the more difficult items of this test and 21 other reference tests were analysed among 640 R.A.F. recruits—a group somewhat superior to the general 18-year male population. The test's homogeneity, as shown by the variances of general and bipolar item factors, is low in this group, but would be just adequate in a representative population. Items of medium difficulty tend to be superior to easy and to very difficult items.

7. Owing to the imposition of a time limit there is a prominent 'difficulty' factor contrasting the earlier and later items. Other formal factors appear to include a general factor in all items (corresponding to the test's specificity in an ordinary analysis of tests), and separate group factors in most of the main sets of which the test is composed. The variance of each of these three kinds of factor in the average item is estimated at close to 5 per cent.

8. The functional content of the items is analysed into *g* 18·4 per cent. and group factors 8·6 per cent., but the figure for *g* might be more than doubled in a representative group. Different items depend to different extents on *v*, *n*, *k*, technical, and perceptual speed group factors; but *k* is much the most important. *V* appears

---

[1] This research was completed before I knew that Miss Keir had carried out a somewhat similar investigation among Secondary Modern Schoolchildren (G. Keir, 'The Progressive Matrices as Applied to School Children,' this *Journal*, 1949, II, 140–150). The results of the two studies generally confirm one another. Thus the order of difficulty of items among adult recruits gives a rank correlation of ·94 with that among Keir's boys. Nevertheless some of the misplacements of items to which she draws attention are reduced among adults; for example the mean difficulty of *D* items becomes greater than that of *C* items. Although the general factor-saturations of particular items differ considerably at the child and adult levels, my second and fourth bipolars sub-divide the items into the same groups as Keir's second and third. In both researches the reliability, general factor-saturations and correlations with other tests are reduced by the restricted range of ability of the testees. But my correlations with eleven verbal spelling and arithmetic tests are somewhat higher, averaging ·63 (range ·54 to ·73), as compared with her average of ·44. And the internal consistency of the test, as shown by the Kuder-Richardson Formula 20, is ·85 as compared with the odd-even coefficient of ·76 among children.

to enter chiefly into the most difficult, and perceptual factor into the easiest, items. Early items are also less *g*-saturated than later ones. But on the whole the test does measure much the same combination of functions at all levels (within the range covered by this group of subjects).

9. If, in a test given with a time limit, the formal or functional content does vary significantly, it is essential that the items be thoroughly mixed (e.g., in cycle-omnibus pattern) and presented in approximate order of difficulty, otherwise different levels of scores will measure different combinations of abilities.

REFERENCES

1. Burt, C. (1939). 'The Latest Revision of the Binet Intelligence Tests.' *Eugen. Rev.*, XXX, 255–260.
2. Burt, C., and Banks, C. (1947). 'A Factor Analysis of Body Measurements for British Adult Males.' *Ann. Eugen.*, XIII, 238–256.
3. Burt, C., and John, E. (1942). 'A Factorial Analysis of Terman Binet Tests.' *Brit. J. Educ. Psychol.*, XII, 117–127, 156–161.
4. Davidson, W. M., and Carroll, J. B. (1945). 'Speed and Level Components in Time-Limit Scores : A Factor Analysis.' *Educ. Psychol. Measmt.*, V, 411–428.
5. Eysenck, H. J., and Crown, S. (1949). 'An Experimental Study in Opinion-Attitude Methodology.' *Int. J. Opinion & Attitude Res.*, III, 47–86.
6. Ferguson, G. A. (1941). 'The Factorial Interpretation of Test Difficulty.' *Psychometrika*, VI, 323–329.
7. Guilford, J. P. (1940). 'Human Abilities.' *Psychol. Rev.*, XLVII, 367–394.
8. Guilford, J. P. (1941). 'The Difficulty of a Test and its Factor Composition.' *Psychometrika*, VI, 67–77.
9. Guilford, J. P., and Lacey, J. I. (1947). *Printed Classification Tests.* Army Air Forces Aviat. Psychol. Prog. Res. Rep. No. 5. Washington, D.C. : U.S. Government Printing Office. Pp. 919.
10. Hertzman, M. (1936). 'The Effects of the Relative Difficulty of Mental Tests on Patterns of Mental Organization.' *Arch. Psychol.*, No. 197. Pp. 69.
11. Kelley, T. L. (1923). *Statistical Method.* New York : Macmillan. Pp. 390.
12. McNemar, Q. (1942). *The Revision of the Stanford-Binet Scale.* Boston, Mass. : Houghton Mifflin. Pp. 189.
13. Raven, J. C. (1939). 'The R.E.C.I. Series of Perceptual Tests : An Experimental Survey.' *Brit. J. Med. Psychol.*, XVIII, 16–34.
14. Rimoldi, H. J. A. (1948). 'Study of Some Factors Related to Intelligence.' *Psychometrika*, XIII, 27–46.
15. Shaefer, W. C. (1940). 'The Relation of Test Difficulty and Factorial Composition Determined from Individual and Group Forms of Primary Mental Abilities Tests.' *Psychometrika*, V, 316–317.
16. Slater, P. (1938). 'Speed of Work in Intelligence Tests.' *Brit. J. Psych.*, XXIX, 55–68.
17. Slater, P. (1948). 'Comment on "The Comparative Assessment of Intellectual Ability".' *Brit. J. Psych.*, XXXIX, 20–21.
18. Smith, G. M. (1933). 'Group Factors in Mental Tests Similar in Material or in Structure.' *Arch. Psychol.*, No. 156. Pp. 56.
19. Spearman, C. E. (1946). 'Theory of General Factor.' *Brit. J. Psych.*, XXXVI, 117–131.
20. Tate, M. W. (1948). 'Individual Differences in Speed of Response in Mental Test Materials of Varying Degrees of Difficulty.' *Educ. Psychol. Measmt.*, VIII, 353–374.
21. Thurstone, L. L. (1938). 'Primary Mental Abilities.' *Psychometr. Monogr.*, No. 1. Pp. 121.
22. Thurstone, L. L. (1938). 'The Perceptual Factor.' *Psychometrika*, III, 1–17.
23. Vernon, P. E. (1938). 'Intelligence Test Sophistication.' *Brit. J. Educ. Psychol.*, VIII, 237–244.
24. Vernon, P. E. (1947). 'Psychological Tests in the Royal Navy, Army and A.T.S.' *Occup. Psychol.*, XXI, 53–74.
25. Vernon, P. E. (1948). 'Indices of Item Consistency and Validity.' *Brit. J. Psych. Stat. Sec.*, I, 152–166.
26. Vernon, P. E. (1949). 'The Structure of Practical Abilities.' *Occup. Psychol.*, XXIII, 81–96.
27. Vernon, P. E., and Parry, J. B. (1949). *Personnel Selection in the British Forces.* London : University of London Press. Pp. 324.
28. Wherry, R. J., and Gaylord, R. H. (1944). 'Factor Patterns of Test Items and Tests as a Function of the Correlation Coefficient : Content, Difficulty, and Constant Error Factors.' *Psychometrika*, IX, 237–244.
29. Wright, R. E. (1939). 'A Factor Analysis of the Original Stanford-Binet Scale.' *Psychometrika*, IV, 209–220.
30. Wrigley, C. F. (1950). *A Psychological Study of Methods of Predicting Success in Flying Training.* Ph.D. Thesis, University of London.

A METHOD FOR COMPUTING PRINCIPAL AXES

By L. F. RICHARDSON

I. *General Explanation.* II. *An Example Begun.*

I. GENERAL EXPLANATION

The following paper describes a working method for calculating the principal components of a correlation or covariance matrix. For reasons which will presently be obvious the procedure adopted has been called 'purification.' A full account of the principles involved is published in the Royal Society's *Transactions* (Richardson, 1950). The present note is intended as a brief guide to the special case for which the given $n \times n$ matrix K is symmetrical. Let the latent roots λ of K be arranged so that $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3 \ldots \geqslant \lambda_n$; and let them correspond respectively to latent columns $P_1, P_2, \ldots, P_n$ by way of the definition

$$KP_j = \lambda_j P_j \qquad (j = 1, 2, \ldots, n) .$$

The method of purification is a generalization of a routine which is familiar to psychologists, and which they usually trace back to Hotelling's paper of 1933; that is to say, the computer begins with a guess $X_\circ$ at a latent column, and he proceeds by iteration. In Hotelling's method $X_\circ$ is multiplied repeatedly by K; and thus P_1 is obtained. My generalization consists in multiplication by $K - \rho I$, where ρ is a guess at an unwanted latent root and I is the $n \times n$ matrix having units in its principal diagonal, and zeros elsewhere. The presence of ρ gives the computer freedom to manœuvre. For example, if he desires to remove P_j, and if he can put $\rho = \lambda_j$, then he can compute $(K - \lambda_j I)X_\circ$, and this does not involve P_j. Chemical language gives suitable hints. Thus $X_\circ$ can be regarded as a raw material containing a desired constituent, which might on occasion be P_n, together with a lot of miscellaneous dirt consisting of $P_1, P_2, \ldots, P_{n-1}$. The multiplier $(K - \lambda_1 I)(K - \lambda_2 I) \ldots (K - \lambda_{n-1} I)$ would remove all that dirt.

To test whether a column Z is a pure latent column, let each element of KZ be divided by the corresponding element of Z, and denote the n quotients by $KZ \div Z$. If the n elements of $KZ \div Z$ are equal to one another, then they are equal to a latent root, and Z is its latent column. This 'division-test' is already standard practice. The only novelty is the notation $KZ \div Z$. There is no need to consider the convergence of infinite sequences.

The iterations are shortened if the raw material $X_\circ$ is rich in the desired constituent P_j. It is customary to start from $\{ 1, 1, 1, 1, \ldots, 1 \}$. But this, though rich in some constituents, is poor in others. It is preferable to choose as the raw material the appropriate latent column of some matrix which resembles the matrix in question.

Rayleigh's mean is a notable detergent. If a column X consists mainly of the latent column P_j, with only slight contamination by the $n - 1$ other P's, then Rayleigh's mean $\tilde{X}KX / \tilde{X}X$ is often an excellent approximation to λ_j. Here the symbol $\sim$ is used to denote transposition.

A precept of the purification method is that the wise computer locates all the latent roots roughly, before proceeding to find any latent root accurately. The iterations begin with the 'initial groping,' which is a game of skill, and follow on with the 'fine purification,' which is a routine. For any one latent root, the stage of groping ends, and fine purification can begin, as soon as all the *other* latent roots have been roughly located.

Some well-known facts may here be usefully recalled. There is a quadratic form in the normalized test scores which represents an n-dimensional surface over which the frequency is constant. The matrix of that quadratic form is the reciprocal of the matrix of correlations (as was shown by K. Pearson, 1896). It follows that the semi-axes of the hyperellipsoid of constant frequency are $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \sqrt{\lambda_3}, \ldots, \sqrt{\lambda_n}$. We may accordingly expect that the least latent root λ_n cannot be negative; for, if it were, then the frequency would be finite in some directions at infinity. This is, however, only an expectation, not a definite proof, for the following assumptions have intervened: (i) that the persons tested were infinitely numerous; (ii) that their distribution was Gaussian-normal; (iii) that the correlations were not rounded off.

Much interest attaches to $\sqrt{(\lambda_n/\lambda_1)}$; because, if this quantity is practically zero, then the hyperellipsoid goes flat; so that $n - 1$ suitably chosen co-ordinates suffice to describe the distribution.

II. AN EXAMPLE BEGUN

For illustration let us take Burt's (1915) study of the correlations between assessments for eight emotional traits. Let us call the matrix of observed correlations K.

TABLE I. CORRELATIONS BETWEEN EMOTIONAL TRAITS (BURT)

$$K = \begin{bmatrix} 1{\cdot}00 & 0{\cdot}83 & 0{\cdot}81 & 0{\cdot}80 & 0{\cdot}71 & 0{\cdot}54 & 0{\cdot}53 & 0{\cdot}24 \\ 0{\cdot}83 & 1{\cdot}00 & 0{\cdot}87 & 0{\cdot}62 & 0{\cdot}59 & 0{\cdot}58 & 0{\cdot}44 & 0{\cdot}45 \\ 0{\cdot}81 & 0{\cdot}87 & 1{\cdot}00 & 0{\cdot}63 & 0{\cdot}37 & 0{\cdot}30 & 0{\cdot}12 & 0{\cdot}33 \\ 0{\cdot}80 & 0{\cdot}62 & 0{\cdot}63 & 1{\cdot}00 & 0{\cdot}49 & 0{\cdot}30 & 0{\cdot}28 & 0{\cdot}29 \\ 0{\cdot}71 & 0{\cdot}59 & 0{\cdot}37 & 0{\cdot}49 & 1{\cdot}00 & 0{\cdot}34 & 0{\cdot}55 & 0{\cdot}19 \\ 0{\cdot}54 & 0{\cdot}58 & 0{\cdot}30 & 0{\cdot}30 & 0{\cdot}34 & 1{\cdot}00 & 0{\cdot}38 & 0{\cdot}21 \\ 0{\cdot}53 & 0{\cdot}44 & 0{\cdot}12 & 0{\cdot}28 & 0{\cdot}55 & 0{\cdot}38 & 1{\cdot}00 & 0{\cdot}10 \\ 0{\cdot}24 & 0{\cdot}45 & 0{\cdot}33 & 0{\cdot}29 & 0{\cdot}19 & 0{\cdot}21 & 0{\cdot}10 & 1{\cdot}00 \end{bmatrix}$$

This matrix, K, has also been studied by Holzinger and Harman (1948, p. 176). My table differs from theirs only by the fact that I have put units in the principal diagonal.

Raw material is needed for purification. The experienced reader may have in stock all the latent columns of some arithmetical matrix which closely resembles the given K. If so, he would do well to take them as his raw material. Here I shall have to be content with a model which has only a distant resemblance to K, but which is so simple that it can be analysed thoroughly by algebra. It is the $n \times n$ matrix A, called A for 'average,' and defined to have units on the principal diagonal, and to have the same number α in each other site.

A standard method for analysing A is to post-multiply it by the column T, called T for 'trial,' and defined thus:

$$T = \{ 1, \omega, \omega^2, \omega^3, \omega^4, \ldots, \omega^{n-1} \}.$$

The curly bracket is used when what is really a column is set horizontal. It is then found that each of the n elements of $AT \div T$ is equal to $1 + \alpha(\omega + \omega^2 + \omega^3 + \omega^4 + \ldots + \omega^{n-1})$ if only $\omega^n = 1$. In order to provide latent columns in sufficient variety, ω is at first taken to be a complex number, and at the end of the theory the real and imaginary parts of T are separated. By this method the n latent columns come out automatically as a mutually orthogonal set. There is a single isolated root, $\lambda_1 = 1 + (n - 1)\alpha$, for which $T = \{ 1, 1, 1, 1, \ldots, 1 \}$. The other $n - 1$ latent roots are each equal to $1 - \alpha$. For the special case $n = 8$ the latent columns are shown in Table II. They do not involve α, and so may be applicable to other 8-rowed matrices of correlations.

TABLE II. ALL THE LATENT COLUMNS OF THE AVERAGED MATRIX A

$$\begin{array}{lcccccccc}
T_0 = \{ & 1, & 1, & 1, & 1, & 1, & 1, & 1, & 1 \;\} \\
T_1 = \{ & \sqrt{2}, & 1, & 0, & -1, & -\sqrt{2}, & -1, & 0, & 1 \;\} \\
T_1' = \{ & 0, & 1, & \sqrt{2}, & 1, & 0, & -1, & -\sqrt{2}, & -1 \;\} \\
T_2 = \{ & 1, & 0, & -1, & 0, & 1, & 0, & -1, & 0 \;\} \\
T_2' = \{ & 0, & 1, & 0, & -1, & 0, & 1, & 0, & -1 \;\} \\
T_3 = \{ & \sqrt{2}, & -1, & 0, & 1, & -\sqrt{2}, & 1, & 0, & -1 \;\} \\
T_3' = \{ & 0, & 1, & -\sqrt{2}, & 1, & 0, & -1, & \sqrt{2}, & -1 \;\} \\
T_4 = \{ & 1, & -1, & 1, & -1, & 1, & -1, & 1, & -1 \;\}
\end{array}$$

The above eight columns may be regarded as jagged waves having respectively the following numbers of crests in the run of eight elements : 0, 1, 1, 2, 2, 3, 3, 4. The suffix to T is the number of crests. Waves having the same number of crests differ only in phase ; and one of them is distinguished by an accent.

There is a preliminary expectation that the latent columns of the given matrix K may resemble those of the matrix A ; and that the resemblance may extend also to the latent roots, if in A we put α equal to the average of those elements of K which are not on the principal diagonal. Then $\alpha = 0{\cdot}4603$. For this special form of A we have one latent root equal to $1 + 7\alpha = 4{\cdot}22$, and seven roots each equal to $1 - \alpha = 0{\cdot}54$.

The raw material is now ready for testing. Each of the columns T belonging to A is to be pre-multiplied by K. The operation is easy because the elements of the T are simple. First,

$$KT_0 = \{\, 5{\cdot}46,\ 5{\cdot}38,\ 4{\cdot}43,\ 4{\cdot}41,\ 4{\cdot}24,\ 3{\cdot}65,\ 3{\cdot}40,\ 2{\cdot}81 \,\}\,.$$

These numbers are also, by exception, equal to $KT_0 \div T_0$. They are not equal to one another ; so T_0 is not a latent column of K. Yet they are of the same order of magnitude. Rayleigh's mean $\widetilde{T}_0KT_0/\widetilde{T}_0T_0$ is, by exception, equal to the simple mean, and comes to 4·22. It may be regarded as a first approximation to λ_1. Next

$$\begin{array}{ll}
KT_1 & = \{\, 0{\cdot}14,\ 0{\cdot}59,\ 0{\cdot}90,\ \ 0{\cdot}05,\ -0{\cdot}46,\ -0{\cdot}23,\ -0{\cdot}15,\ 1{\cdot}02 \,\} \\
KT_1 \div T_1 & = \{\, 0{\cdot}10,\ 0{\cdot}59,\ \infty,\ -0{\cdot}05,\ \ 0{\cdot}33,\ \ 0{\cdot}23,\ \ \infty,\ \ 1{\cdot}02 \,\}
\end{array}$$

These ratios are alarmingly unequal. The infinities arise, however, only from zero elements in the divisor. In Rayleigh's mean the zero elements act, not as divisors, but as multipliers ; and so they do no harm. It is

$$\widetilde{T}_1KT_1 \,/\, \widetilde{T}_1T_1 = 2{\cdot}64/8 = 0{\cdot}33\ .$$

The same method has been applied to the columns T_1', T_2, T_2', T_3, T_3', T_4. The estimates which Rayleigh's mean gives for the eight latent roots of K are

$$\begin{array}{lcccccccc}
\lambda = & 4{\cdot}22 & 0{\cdot}96 & 0{\cdot}64 & 0{\cdot}55 & 0{\cdot}52 & 0{\cdot}48 & 0{\cdot}33 & 0{\cdot}28 \\
\text{From} & T_0 & T_1' & T_2' & T_4 & T_3' & T_3 & T_1 & T_2
\end{array}$$

The sum of the eight estimates is 7·98. The sum of the true λ is equal to the sum of the principal diagonal of K, namely, 8 exactly. The agreement is close, but must not be allowed to engender too much confidence in these preliminary estimates ; for none of the division-tests was anywhere near satisfied. The seven equal roots of the averaged matrix A appear to have spread out a bit now that A has been replaced by K ; but the mean of those seven roots remains unaltered at 0·54. Altogether the averaged matrix A seems to be justified as a useful beginning. The raw columns T can now be purified with the aid of the preliminary λ which they have provided.

To find the greatest root λ_1. The other roots appear to form a cluster, well separated from λ_1. This allows us to ' kill seven birds with one stone,' namely, to remove most of the $P_2, P_3, P_4, P_5, P_6, P_7, P_8$, which occur as impurities in T_0, by way of the single multiplier $K - 0{\cdot}54I$. Thus let P_1^2 denote a second approximation to P_1 defined by

$$P_1^2 = KT_0 - 0{\cdot}54\, T_0 \quad .$$

It is important to notice that the upper of the two affixes to P is not the index of a power, but is the serial number of an approximation.

It follows that

$$P_1^2 = \{\ 4{\cdot}92,\ 4{\cdot}84,\ 3{\cdot}89,\ 3{\cdot}87,\ 3{\cdot}70,\ 3{\cdot}11,\ 2{\cdot}86, 2{\cdot}27\ \}$$
$$KP_1^2 = \{\ 21{\cdot}55, 20{\cdot}97, 17{\cdot}92, 17{\cdot}46, 16{\cdot}45, 13{\cdot}72, 12{\cdot}59, 9{\cdot}68\ \}$$
$$KP_1^2 \div P_1^2 = \{\ 4{\cdot}38,\ 4{\cdot}33,\ 4{\cdot}61,\ 4{\cdot}51,\ 4{\cdot}44,\ 4{\cdot}41,\ 4{\cdot}40, 4{\cdot}26\ \}$$

The elements of $KP_1^2 \div P_1^2$ are much more alike than were those of $KT_{\circ} \div T_{\circ}$. Considerable purification has evidently occurred. Rayleigh's mean is

$$\widetilde{P}_1^2 K P_1^2 \big/ \widetilde{P}_1^2 P_1^2 = 4{\cdot}425,$$

and is probably correct to the first or second decimal. If further accuracy is desired as to λ_1 and P_1, it can rapidly be attained by repeated multiplication by $K - 0{\cdot}54I$. However, as this paper is about method rather than results, I leave the routine for refining λ_1 aside as obvious, and turn instead to :—

Approximations to the least latent root λ_8 *and its latent column* P_8. The column T_2, which yielded the least estimate of λ, has been chosen as the raw material. The choice of ρ, in the multiplier $K - \rho I$, is discussed in the full account of the method. For reliability everywhere ρ must exceed $\frac{1}{2}(\lambda_1 + \lambda_8)$. If nothing is known about the positions of the intermediate roots, $\lambda_2 \ldots \lambda_7$, then the best choice of ρ, for a single multiplication, is $\rho = \frac{1}{3}(2\lambda_1 + \lambda_8)$, which comes to about 3·0. Repeated multiplication of T_2 by $K - 3I$ would undoubtedly yield P_8; but the process would be insufferably slow, because of the cluster of roots close to λ_8. It is better here to take advantage of the special information that there is no latent root between 1 and 4·2. In this gap the purifier need not be reliable. Two multipliers can be used in conjunction, one to remove P_1, the other to remove $P_2, P_3, P_4, P_5, P_6, P_7$. The multiplier $(K - 4{\cdot}43I)$ will remove P_1 neatly, and is reliable everywhere. For the purpose of removing P_2, P_3, P_4, P_5, P_6, P_7, the constant ρ must exceed $\frac{1}{2}(\lambda_2 + \lambda_8)$. For a single multiplication by $K - \rho I$ it would be suitable to take $\rho = \frac{1}{3}(2\lambda_2 + \lambda_8)$, which comes to 0·73 on the uncertain information available at this stage. Actually I began with $\rho = 0{\cdot}7$ and afterwards decreased ρ to 0·6 when it became clear that λ_8 was less than expected. The multiplier $(K - 0{\cdot}7I)$ is not reliable everywhere : it both cleans and dirties. When the amplitude of P_8 is taken as the standard, then the multiplier $(K - 0{\cdot}7I)$ diminishes the amplitudes of P_7, P_6, P_5, P_4, P_3, P_2, but greatly magnifies the amplitude of P_1. To repress P_1 multiplication by $(K - 4{\cdot}43I)$ is used. The combined multiplier $(K - 4{\cdot}43I)(K - 0{\cdot}7I)$ is reliable everywhere, except in the gap where there are no latent roots. If the computer desires to multiply by a high power of a matrix, obtained by repeated squaring after the manner of Hotelling (Holzinger and Harman, 1948, p. 166), it would be suitable for him to work out a high power of $(K - 4{\cdot}43I)(K - 0{\cdot}7I)$. Here, however, in order to illustrate the process, I have used multipliers that are linear in K. It is important to notice that the impurity P_1 cannot be removed once for all at the outset. This is impossible partly because 4·43 is an inaccurate estimate of λ_1. But even if λ_1 were known precisely, nevertheless small amplitudes of P_1 would creep in whenever the digits were curtailed, and these traces of P_1 would be magnified in the subsequent multiplications by $(K - 0{\cdot}7I)$. When $(K - 0{\cdot}7I)$ and $(K - 4{\cdot}43I)$ are used alternately, the impurity P_1 alternately increases and diminishes, and may hide the steady decrease in the other impurities. In these circumstances the computer needs to make sure that there are no mistakes in his arithmetic ; he needs also to understand the theory of purification, and to have faith in it. Table III shows that the first multiplication by $(K - 0{\cdot}7I)$ did harm as well as good, for it increased the estimate of λ_8 obtained by Rayleigh's mean. A subsequent multiplication by $(K - 0{\cdot}6I)$ had a similar effect.

TABLE III. SUCCESSIVE APPROXIMATIONS TO THE LEAST LATENT ROOT OF K

Estimates of Latent Column P_8		Estimates of Latent Root λ_8		
		By Division-test		
How Obtained	Symbol	Least	Greatest	By Rayleigh's Mean
From matrix A	T_2	$-0{\cdot}06$	∞	0·28
$KT_2 - 0{\cdot}7T_2$	P_8^1	$-3{\cdot}6$	∞	1·13
$KP_8^1 - 4{\cdot}43P_8^1$	P_8^2	$-1{\cdot}11$	0·91	0·13
$KP_8^2 - 0{\cdot}7P_8^2$	P_8^3	$-2{\cdot}52$	0·054	0·11
$KP_8^3 - 0{\cdot}7P_8^3$	P_8^4	$-0{\cdot}174$	0·212	$-0{\cdot}005$
$KP_8^4 - 0{\cdot}6P_8^4$	P_8^5	$-1{\cdot}273$	1·978	0·031
$KP_8^5 - 4{\cdot}43\,P_8^5$	P_8^6	$-0{\cdot}296$	0·029	$-0{\cdot}012$
$KP_8^6 - 0{\cdot}6\,P_8^6$	P_8^7	$-0{\cdot}038$	0·310	$-0{\cdot}013$
$KP_8^7 - 4{\cdot}43\,P_8^7$	P_8^8	$-0{\cdot}036$	0·274	$-0{\cdot}017$

I have no doubt that λ_8 is negative, and that it can be excused as practically zero. The last approximation to P_8 is

$$P_8^8 = \{\,92{\cdot}5669,\ 49{\cdot}2177,\ -78{\cdot}3811,\ -26{\cdot}0161,\ -28{\cdot}1690,\ -26{\cdot}9881,\ -29{\cdot}8714,\ 4{\cdot}2068\,\}$$

To find the second latent root it would be suitable to multiply T_1' repeatedly by $(K - 4{\cdot}43I)(K - 0{\cdot}3I)$.

But perhaps enough has been done to illustrate the method of purification.

I am obliged to Sir Cyril Burt for various guidance.

REFERENCES

1. Burt, C. (1915). *Report of the British Association for the Advancement of Science*, pp. 694–696.
2. Holzinger, K. J., and Harman, H. H. (1948). *Factor Analysis*. Chicago: University of Chicago Press.
3. Pearson, Karl (1896). *Roy. Soc. Lond. Phil. Trans.*, A, CLXXXVII, 301.
4. Richardson, L. F. (1950), *Roy. Soc. Lond. Phil. Trans.*, A, CCXLII, 439–491.

NOTE.—At Dr. Richardson's request I append the figures I myself obtained for the factor-saturations (latent vectors before normalization) and factor variances (latent roots) for the two largest factors. These results may be of assistance to any reader who desires to try his own hand at Dr. Richardson's instructive method of calculation. The figures given below were obtained to suit the requirements of Karl Pearson's "method of principal axes"; they were computed by weighted summation with unity in the leading diagonal (as in Table I above).

Trait	1	2	3	4	5	6	7	8	*Latent Root*
Factor I	·961	·933	·800	·780	·731	·608	·557	·427	4·439
Factor II	·047	−·145	−·452	−·186	·369	·244	·677	−·464	1·132

C. Burt

A STATISTICAL STUDY OF THE RORSCHACH TEST

By AMYA SEN

Department of Psychology, University College, London

I. *Problem.* II. *Subjects and Tests.* III. *The Validity of the Rorschach Results: (a) Cognitive Traits; (b) Orectic Traits.* IV. *A Factor Analysis of Rorschach Categories.* V. *Summary and Conclusions.*

I. PROBLEM

The Validity of the Rorschach Tests. The object of the investigation described in the following paper was twofold: first, to compare the validity of the results obtained with the Rorschach inkblots (*a*) when the test is scored according to one of the recognized Rorschach methods, and (*b*) when it is scored in accordance with the principles suggested by Professor Burt (10); and secondly to determine objectively, by means of a factorial analysis, the most appropriate classification of the various characteristics revealed by the test, and to compare this objective classification with the classifications of personality types which the test is said to reveal when interpreted subjectively by an orthodox Rorschach technique.

The inkblot test as standardized by Rorschach and his followers has been described as "the most widely used tool in diagnostic personality testing" (7, p. 85). Yet it has been strongly criticized by those whose work lies specifically in the field of test validation. Burt, for example, states that, "as ordinarily scored, it yields validity coefficients that are far lower than assessments for the same traits based on an interview and lasting for about the same length of time" (10). Cattell declares that "despite the attraction which the Rorschach test has exercised, especially on psychiatrists, it remains a mixture of ill-defined intentions, analogous to a patent medicine, devoid of clear-cut theoretical bases" (21). Eysenck is equally trenchant: "To many Rorschach experts this test seems to fulfil the functions which in physics would fall to a spectroscope, cyclotron, thermionic valve, and closed chamber all in one: to more cautious psychologists, it appears as a test whose reliability is low, whose validity has never been established, and whose subjective nature does not attract the scientific worker" (22, p. 213). And Goodenough, after pointing out that the 'formulæ' for interpreting the scores seem largely to have been "developed on the basis of chance relationships," concludes, a little more cautiously, that "the Rorschach method has now reached a point where a more critical attitude is desirable" (25, p. 436).

Empirical Validation. In general the critics of the test lay stress on two main defects. In the first place, despite its widespread use, hardly any researches have been undertaken to measure its reliability and validity by recognized statistical devices. There are plenty of investigations showing that it 'discriminates' between various groups—particularly between the neurotic and the non-neurotic, and between different types of neurotics. But a test may yield unquestionable evidence for 'discrimination' (i.e., for a fully significant difference between the group-averages), and yet possess little or no value as a diagnostic guide. As Rapaport observes, "the published comparisons of blind diagnoses with clinical histories and symptomatology do not validate the test: they merely demonstrate its potentialities" (20, p. 86; cf. also Goodenough, 25, p. 436).

PSYL. RESEARCH LIBRARY
6.8.70

One of the objects of the following paper will therefore be to provide quantitative indications as to the *amount* of validity that attaches both to the separate categories and to the test as a whole.[1]

Theoretical Construction. In the second place, as Rapaport points out, "the tie-up of the test with theoretical development in general psychology and the psychology of personality has also lagged." The traits or characteristics which the Rorschach test claims to assess consist in the main of an improvised list of psychological qualities which Rorschach himself suggested nearly thirty years ago ; and these for the most part assume the popular but highly fallacious principles so often stigmatized in criticisms of the faculty doctrine as the 'naming fallacy' and the 'formal fallacy.' In consequence, as even the Rorschach expert himself at times admits, there is considerable danger that "the tester, especially the novice," may be led to a "dream-book type of interpretation" (20, p. 88). Goodenough makes a similar complaint. "There is no obvious resemblance between the nature of the subject's responses and the personality traits from which they are alleged to spring. Many of the 'signs' reported have stemmed for the most part from 'hunches'" (25, p. 435).

Nevertheless, 'the scientific worker' is scarcely justified in dismissing a widely used test outright, without at least attempting to discover what degrees of validity its assessments actually possess, or suggesting constructive proposals for its modification and improvement. With the view, therefore, to making some systematic contribution towards these two main problems, a number of researches have been planned and carried out in the laboratory of University College. Of these, the earliest, so far as the Rorschach test itself is concerned,[2] was that undertaken by Miss Kerr, who applied the test to 265 normal, maladjusted, and defective children (11). A number of recent theses, published and unpublished, based on children or more frequently on adults, have dealt either with the Rorschach test as such, or with the Rorschach test as one of several methods of assessing personality (cf. 17).[3]

II. SUBJECTS AND TESTS

Subjects. One common line of criticism has been to suggest that the alleged correspondences between test scores and personal characteristics might be of social rather than of psychological origin. It seemed therefore desirable to obtain results from persons brought up under somewhat different cultural conditions from those who have generally been the subjects in most of the researches hitherto published.

---

[1] The few investigations that have attempted to assess *validity* (e.g., by correlating Rorschach scores with an independent criterion) have been concerned almost exclusively with the assessment of intelligence. The chief exception is that of Dr. P. E. Vernon (12, 13). In Table I of my thesis (27) I have compiled a comparative summary of the figures obtained by different workers, and have shown that they are extremely conflicting.

The *reliability* of the scoring and of the test-results was not a primary object of the present research, though a small sample of the scripts was scored independently by Ramzy. The results are given later in this paper. Miss Hillman has demonstrated that the reliability of the different blots, judged by two sets of scores based on half the blots, yields only a moderately high tetrachoric correlation (0·60 to 0·75 : unpublished thesis). In regard to the reliability of different examiners, Ramzy and Pickard have shown that, provided the responses are marked by two psychologists who have been in personal consultation over general principles, the agreement is reasonably close (24).

[2] In several earlier investigations an inkblot test had been used in London schools. In these cases, however, a composite battery of tests was generally employed, similar to that used by Burt and Moore, in which the test-material included not only inkblots and spot-patterns, but also ambiguous or indeterminate pictures, unfinished drawings and stories, as well as tests of intelligence. Of these researches the most suggestive was that of H. L. Hargreaves on 'The Faculty of Imagination' (*Brit. J. Psych. Mon. Sup.*, X, 1927) who used tests of this nature with 131 children in L.C.C. schools. 'Inkblots' and 'indeterminate pictures' both proved to be valuable tests for two aspects of imagination termed 'fluency' and 'originality.'

[3] I should like to acknowledge my indebtedness to all those who have assisted me in my own investigation, and particularly to Dr. Charlotte Banks, who has given special help both in the statistical analysis and in the writing of my paper.

Accordingly, for the present study, 100 Indian students were selected ; these comprised 60 men and 40 women, aged for the most part between 20 and 25, working at various subjects either in London or in Cambridge.

Tests. Each subject was given an individual Rorschach test, two intelligence tests (Group Test No. 33, a verbal group test prepared by Burt for the National Institute of Industrial Psychology, and the Progressive Matrices test, a non-verbal group test prepared by Raven), and Cattell's test[1] of fluency (based largely on Hargreaves' work, cited above). At a later stage they also took two questionnaires intended to elicit relevant traits of personality (for details see 27).

Scoring of the Rorschach Test. (*a*) *Rorschach Method.* Rapaport has suggested that "the lines along which the examiner should organize the scores" may be described as the "dimensions" of the tests, and that, in marking responses, these "dimensions" should be treated "as though they were a system of co-ordinates within which the individual case can be located." The term itself is intended to imply that all the response-characteristics grouped within one 'dimension' are "pertinent to some major aspect of the subject's psychological make-up, and independent of other major aspects" (20, p. 109). However, the classification of response-characteristics for these purposes is evidently based, not on any factorial analysis of the test-results, but on impressionistic ideas of the various "psychological functions of the subject" which the test is supposed to reveal.

The "general rationale of the psychological processes underlying Rorschach responses," we are told, depends on the hypothesis that "the examiner may see in the subject's reactions to the inkblots perceptual organizing processes which are fundamentally continuous with his perceptions in everyday life." If, for example, the subject is observant when looking at inkblots, then, it is assumed, he will be observant in ordinary life ; if his responses are largely concerned with the dark shadows in the blots, then, it is supposed, he will tend to be unduly preoccupied with the dark shadows in his own existence ; and so on. The actual content or material of his responses is treated as of little consequence : "only in very sick patients, or cases with acute problems, does the content become directly revealing" (*ibid.*, p. 102). It is the 'form' of the subject's mental processes that constitutes the main basis of the scoring. The general ways in which he reacts in seeking to give some intelligible interpretation to the visual stimuli before him are regarded as a sample of the formal characteristics that distinguish his mental reactions in actual life.

Among German writers on individual psychology this mode of interpretation has always been widely prevalent : it is admirably exemplified in their method of deducing personal characteristics from the corresponding characteristics seen in the person's handwriting. Among English-speaking writers, on the other hand, from Thorndike onwards,[2] the notion of "forms of mental activity, which work alike in every kind of concrete situation," has been highly suspect. It is true that recent factorial studies have at times confirmed the existence of a few formal factors, common to certain types of mental process : but, with rare exceptions, their contribution to the variance has proved to be small—far too small for diagnostic purposes.

For the detailed nature of the various 'categories' commonly adopted by Rorschach workers, and of the scoring procedure generally employed, the reader may be referred to the standard works on the subject (9, 18, 19). Here it will be sufficient to note that the scores are usually classified under four main heads : (i) *Quantity*, i.e., number of responses and time taken ; (ii) *Area* or *Location*, e.g., how far the response covers the entire blot, or only parts, or possibly just minor details, and how successfully the response organizes the main constituents into a logically articulated whole ; (iii) *Determinants*, i.e., references to (*a*) colour, (*b*) form, (*c*) shading, (*d*) movement, and particularly (*e*) the relative balance of each ; (iv) *Content*, i.e., whether the things seen in the tests tend to be original or banal, and particularly what proportion can be assigned to certain stock headings—classes which are regarded as more or less symbolic, human beings, other animals, blood, fire, mountains, caves, sexual objects, and the like, and whether they show marked repetition or perseveration : it will be noted that, even in assessing 'contents,' it is still the broad formal characteristics that are noted, not their specific nature or their direct implications.

In the following enquiry, in order to obtain scores according to approved Rorschach principles, Beck's procedure (18) was used, with two minor changes. A slight modification of his '*Z*-score' was introduced along lines suggested by the present writer,[1] and Klopfer's 'form level score' (19) was used in addition, with a view to determining which procedure was most effective. As a check on the reliability of the scoring, Mr. Ramzy was good enough to score 25 of the scripts independently: the contingency coefficients for 'location' and the various 'determinants' averaged 0·86 and 0·90 respectively.

---

[1] Cattell, R. B. (1948). *A Guide to Mental Testing*, pp. 177f.

[2] Thorndike, E. L. (1923). *J. Educ. Psychol.* 'Individual Differences and their Causes.' III, esp. pp. 365f.

(*b*) *Burt's Method.* The principles underlying Burt's suggestions may be summarized in his own words as follows.

"Inkblots have long been used as test-material, both in experimental and in individual psychology. Before James' criticisms [2] put the term out of fashion, they were regularly employed to exemplify what neo-Herbartian writers called 'apperception.' 'Apperception' may be defined as the mental process whereby the contents of a particular sensory presentation, at first relatively indeterminate and meaningless, become clearer and more determinate, and so acquire for the perceiver some kind of *meaning*, in virtue of a fusion or synthesis with contents already more or less systematically organized in his mind.[3] It is the fashion nowadays to describe the effects of this process by a popular metaphor from the magic-lantern: the perceiver is said to 'project' [4] his ideas and feelings in his mind on to the stimulus he is perceiving—an analogy which easily leads to extremely naïve, crude, and unjustifiable deductions. After all, thanks to the experimental work of Sharp,

---

[1] 27, p. 75. It seems more in accordance with the principles of scoring adopted by Beck and other Rorschach testers to regard the following types of response as not falling within the scope of Beck's *Z*: responses in which a small detail has suggested a comment bearing no genuine relation to the whole blot, responses consisting of combinations of small and large details where these are plainly inaccurate, and finally, vague, indefinite responses, such as 'maps' or the like. All these would receive a *Z*-value according to the official instructions for Beck scoring. In my own investigation both the unmodified and the modified *Z*-score were calculated and compared; but, as will be seen from the final results, the latter seemed unquestionably preferable.

[2] James, W. *Principles of Psychology*, II, pp. 107f. James quotes a long passage from Steinthal, who has perhaps given the most thorough analysis of the 'apperceptive processes' and their possibilities. Steinthal points out that, in ordinary everyday life, our commonest clues to a man's character and their place in life are furnished by his "mode of apperceiving." James comments that he himself finds Steinthal's analysis 'simply burdensome.' But an analytic study of the different processes involved is surely preferable to lumping them all together as comparable with the simple and direct 'projection' of a slide on to a screen.

[3] Stout, G. F. (1899). *Analytic Psychology*, II, ch. viii; (1913), *Manual of Psychology*, pp. 147f. As examples he cites inkblots, puzzle pictures, and clouds or constellations in the sky. Seashore's definition is similar to Stout's: "apperception designates the fusion of the presentative with the representative processes, including all their affective and conative processes, in the moment of a new experience. It is the grasping of a new experience so as to give it meaning and clearness. The interpretation is always personal" (*loc. cit. inf.*, p. 144). It may perhaps be added that, since the above was written, the term 'apperception' seems to be coming once more into favour: thus Bell commonly translates Rorschach's terms *Erfassung* and *Wahrnehmung* by 'apperception', (e.g., 23, pp. 88 and 190), and describes the 'determinants' as indicating 'quality of apperception.'

[4] Elsewhere Burt points out the widely different kinds of mental process that the ambiguous term 'projection' is made to cover. Bell, Rapaport, and Goodenough, for instance, all treat the Rorschach test as a typical "projective procedure"; but they explain the term in very different ways. Goodenough gives three explanatory examples: (i) "We say our soup is cold": here we attribute the "sensation in our mouths" to the soup outside us; (ii) "the office-manager, during an attack of indigestion, raises a tempest over a misfiled letter; he projects his feelings on to the persons and things surrounding himself." But he does not actually attribute his angry *feelings* to the letter or the clerk. (iii) "In some forms of mental disease projection is carried to the point of delusion: the patient sees demons or hears voices of angels." Here hallucinatory reality is attributed to a visual or auditory image—again a different process. Rapaport, on the other hand, attributes the projection not to the person but to the test: "the concept of projection in projective procedure is formed on the pattern of projector and screen." No wonder Bell observes that "no clear or common definition of what is meant by 'projective' has as yet appeared among those who use these methods." He is quite wrong, however, in stating that "projection was first used in a psychological sense by Freud." The idea of such a process has always been a favourite one with physiological and medical writers; and was criticized long ago by James (*Principles*, II, p. 31). As Ward has put it, "those who talk about 'projection' have usually committed the prior fallacy of 'introjection'."

Pillsbury, Bartlett, and others,[1] we do know a good deal about the varied processes involved in apperceiving inkblots, though writers on the Rorschach tests make little or no mention of it.

"Older books on experimental psychology commonly started with a laboratory experiment of this type to demonstrate to the student how much of what he thinks he sees is really an unconscious contribution of his own mind, and how characteristically the interpretations offered differ from one observer to another. 'To a hunter,' it is said, a particular inkblot 'suggests a beaver or a woodchuck; to a mason, it is a trowel; to a keeper of pets an Angora cat; to the naturalist a hedgehog or flounder, according as his mind has been most directed to land animals or to fish.'[2]

"It follows that, by comparing the particular ways in which different individuals *apperceive* the same series of 'equivocal stimuli,' we may often secure a detailed insight into their habitual vocabulary and cultural background, their dominant interests and ideas, their past history and experiences, and even their emotional attitudes and complexes. In all such cases, however, it is the *contents* of the 'apperceptive contributions,' as disclosed by the contents of their interpretations, that provide us with the requisite data, rather than the *formal* characteristics of their perceptual or associative processes. . . . In summarizing the inferences drawn according to this principle it will be natural to classify the traits in accordance with the usual type of psychographic scheme, elaborated by the workers in the field of individual psychology, instead of inventing some new scheme and some special list of categories, expressly for the purpose of this particular test."

In the present enquiry the assessments were limited to those particular traits which previous trials had shown could be most reliably rated. Accordingly, data extracted from responses were classified under the following heads: Intelligence, Verbal Ability, Imagination, Perception of Relations, General Emotionality, Extraversion-Introversion, Assertion-Submission, Cheerfulness-Depression, Sociability, Anxiety, Neurotic Tendencies, and Extent of Vocational and Cultural Interests. These largely correspond with mental factors already established in the fields of cognitive or orectic psychology. The numerical ratings were obtained by counting up the number of relevant items in accordance with the scheme suggested by Burt; the score for each trait was then converted into terms of a 15-point scale, running from E– to A+.[3]

Criteria. Each of the subjects tested were rated independently by two judges for the traits enumerated above. The judges were friends or close acquaintances of the subjects whom they rated, and had lived under the same roof with them for at least two years. It proved impossible to find a single pair of judges who were well acquainted with each one. Three pairs were able to rate groups of 15, 13, and 10 subjects respectively; the rest rated somewhat smaller groups. Each was asked to assess the subjects in his (or her) group for one trait at a time; and the assessments were made on a 15-point scale, similar to that used for rating the Rorschach scores.

Reliability of Judges. To estimate the reliability of the ratings, an overall correlation was first calculated between the two ratings thus secured for every trait, regardless of any differences in the judges. The coefficients ranged from 0·510 to 0·791, and averaged 0·712. Separate correla-

[1] Pillsbury, W. B. (1896). 'A Study of Apperception.' *Amer. J. Psych.*, VII, pp. 350–393. Sharp, Stella (1899). 'Individual Psychology: a Study in Method.' *Ibid.*, X, pp. 329–391. Bartlett, F. C. (1916). 'An Experimental Study of Some Problems of Perceiving and Imagining.' *Brit. J. Psych.*, VIII, pp. 222–266. Cf. also Hicks, Dawes (1914). 'The Nature and Development of Attention.' *Ibid.*, VI, pp. 7f.

[2] Quoted from Seashore, C. E. (1908), *Elementary Experiments in Psychology*, Ch. XII. Cf., also Witmer (1902), *Analytical Psychology: A Practical Manual*. Ch. I, esp. Experiments III and IV on 'The apperceptional contribution in perceptions of equivocal stimuli' and 'Variable and individual apperception.' In his own experiments on individual differences at Liverpool, Burt used, in addition to material taken from Witmer and Sanford, ambiguous pictures and artificial asymmetrical blots (uncoloured and coloured), which he found in later experiments slightly superior to the symmetrical natural blots of the kind used by Rorschach: both the average reliability (0·41) and the average validity (0·21) were low; but sex differences were easily demonstrated (*J. Exp. Ped.* (1911), I, p. 381; *Brit. J. Educ. Psych.* (1945), XV, pp. 116f.). To control the presentation, and investigate the processes commonly involved, he employed McDougall's 'portable tachistoscope' (which can be adapted to the ordinary lantern for group presentation). Subsequently, as Myers' assistant at Cambridge, he started a more intensive study of 'the general apperceptive processes'; and this later work formed the basis of the present scheme.

[3] The details of the method are described and illustrated in unpublished *Laboratory Notes on the Rorschach Test* (10). For the psychographic schedule employed, see 26 and refs.

tions were then calculated between the two ratings for the three largest groups, each rated by the same pair of judges. The figures obtained ranged from 0·450 to 0·855, and averaged 0·642. The general size and range of the reliability coefficients is thus not very different in the two cases. Since by mixing ratings secured from different judges the size of the coefficients is not significantly increased, we may assume that their standards were pretty much the same. In general, traits which are openly manifested in everyday intercourse, such as intelligence, sociability, intellectual and cultural interests, verbal ability, anxiety, and cheerfulness, were assessed with the highest reliability (about 0·75 to 0·80); traits like imagination or general emotionality were assessed with the lowest reliability (about 0·50 to 0·60).

III. THE VALIDITY OF THE RORSCHACH RESULTS

(*a*) *Cognitive Traits.* (i) *General Intelligence.* To secure a validating criterion for general intelligence, the scores in the two intelligence tests were reduced to standard measure, and averaged. These average scores were then correlated, first, with the judges' ratings for intelligence, secondly, with the ratings for intelligence obtained from the Rorschach responses with Burt's procedure, and thirdly, with certain 'categories,' scored by Beck's method, which are commonly claimed by orthodox Rorschach workers to be indicative of intelligence.

The correlation between the intelligence tests and the judges' ratings was, for the men and women together,[1] 0·671. The correlation between the tests and the ratings given by Burt's content method was 0·765 (men, 0·738; women, 0·792). The correlations between the tests and the several Rorschach categories, scored by Beck's method, are set out below in Table I. In each case the coefficient was calculated by the full product-moment formula.

TABLE I. GENERAL INTELLIGENCE: CORRELATIONS BETWEEN INTELLIGENCE TESTS (CRITERION) AND RORSCHACH CATEGORIES

CATEGORY	CORRELATION	
	Men	Women
Responses to Whole Blot (W%)	—·060	·062
Large Details (D%)	·102	·052
Large Unusual Details ($D + Dd$%)	·046	·056
Good Form ($F +$ %)	·457	·436
Movement Responses (M%)	·408	·415
Beck's Z-score	·221	·403
Modified Beck's Z-score	·409	·415
Klopfer's Form Level Score	·500	·586
Total Number of Responses	·055	·053
Animal Responses and Animal Detail ($A + AD$%)	—·365	—·190
Popular Responses (P%)	·248	—·143

---

[1] Rorschach himself and most other Rorschach workers hold that there are no perceptible differences between the sexes, and therefore do not tabulate results from men and women separately. Burt and Moore, however, found marked sex-differences with their tests, particularly among children; and considered these to be analogous to the different sex interests reflected in other tests of apperception and association. Thus, in experiments carried out during 1914–18, "where the blots or the pictures are coloured, the girls were far more prompt to notice the fact than the boys; on the other hand, the boys made more references to action. Among the boys there were nearly three times as many references to the war as among the girls. . . . Among adults, however, except for occupational and domestic interests, sex-differences were much smaller." Most other workers with inkblots have also reported distinct sex-differences, at least among children of school age (cf. 4 and refs. 8, 11).

In my own groups, perhaps owing to their relatively small numbers, perhaps owing to their different cultural background, sex-differences in the Rorschach scores were in most instances comparatively slight. It therefore seems legitimate to pool the results from the two sexes in computing the correlations. Wherever there is any doubt on this point, I have tabulated coefficients for the two sexes separately.

Coefficients of 0·255 and 0·330 mark the 5 per cent. significance level for the men and the women respectively; the significance level for the average for both sexes is 0·196. It will be seen that the results obtained from the two sexes separately tend to confirm each other. Four categories give correlations over 0·400 for both sexes, namely, the modified Beck score, Klopfer's form level score, percentage of good form, and percentage of movement. None of these, however, approaches the figures obtained by the content method.

(ii) *Other Cognitive Traits.* The judges' ratings for Imagination, Verbal Capacity, and Vocational and Cultural Interests were correlated with the ratings for the same traits obtained by Burt's method; and the judges' ratings for Imagination were correlated with two Rorschach categories claimed to be indicative of this quality, namely, percentage of 'movement' and of 'form' responses. (The Rorschach scoring gives nothing comparable to the ratings for the other two traits.) The coefficients are shown below in Table II. Since here there are no appreciable differences between the figures for the men and women, the results for the two sexes have been pooled into a single group.

TABLE II. COGNITIVE TRAITS: CORRELATIONS BETWEEN JUDGES' RATINGS AND RORSCHACH RESULTS (*a*) WITH BURT'S METHOD, (*b*) WITH BECK'S METHOD

TRAITS	A. Content Method	B. Beck's Method *Percentage of Movement*	*Percentage of Form*
Imagination	·445	·127	−·212
Verbal Ability	·535	—	—
Intellectual Interests ..	·486	—	—

As before, with a group of this size a coefficient of ±0·196 marks the level of statistical significance. For Imagination the content method of rating yields a much higher correlation than the standard procedures. For verbal ability and intellectual interests the results confirm the expectation that an analysis of content will provide a fair indication of these qualities; and since this is such an obvious component of the material obtained, it seems unfortunate to omit it (as the orthodox Rorschach procedure does) from the inferences to be drawn from the test.

TABLE III. ORECTIC TRAITS: CORRELATIONS BETWEEN JUDGES' RATINGS AND RORSCHACH RESULTS (*a*) WITH BURT'S METHOD, (*b*) WITH BECK'S METHOD

TRAIT	CATEGORY	A. *Content Score*	B. *Category Score*
General Emotionality ..	Percentage of Colour Responses ($C\%$)	·600	·198
Extraversion-Introversion	Percentage of Movement minus Colour Responses ($M - C\%$)	·540	·061
Assertion-Submission ..	.. ..	·433	—
Cheerfulness-Depression	Number of Form-Vista plus Flat Grey Responses ($FV - FY$)	·519	·082
Anxiety	Number of Form-Vista plus Flat Grey Responses ($FV - FY$)	·624	·138
Sociability	Number of Human Responses ($H\%$)	·479	·230
Cultural Interests ..	.. ..	·663	—

(*b*) *Orectic Traits.* We now turn to the estimates for affective and conative traits. Once again the ratings of the two judges for each student were averaged; and the averages were then correlated with the Rorschach ratings as obtained by the two alternative procedures. The results are shown in Table III.

The highest correlation with the Rorschach categories is that between the 'number of human responses' and the ratings for sociability. But the low figure (·230) demonstrates that, even in this field, the Rorschach categories do not yield very useful results.

Matching Method. It is sometimes admitted by Rorschach workers that the individual Rorschach categories, taken by themselves, are of but little use for diagnostic purposes. Rather, they say, it is the total 'psychogram,' the *Gestalt*, as it were, of all the categories, that alone will give a successful indication of the real *Struktur* of the individual's personality. To test this claim, the Matching Method provides the most appropriate statistical technique. A verbal character sketch is compiled from each Rorschach protocol; and independent judges are then asked to match the names of the subjects with the several sketches submitted.

In the few researches on the Rorschach test where this method has been used, fairly high coefficients of agreement have commonly been obtained (23, p. 135). But there is some evidence to suggest that this may be due in part to limitations in the method itself. Vernon has pointed out certain drawbacks in the technique that are sometimes overlooked (14). He notes that the results (i) vary greatly with the heterogeneity of the group to be matched; (ii) depend on the experience and insight of the judges; and (iii) demand a good deal of experience and skill on the part of the interpreter who interprets the Rorschach responses and writes the sketches. One further drawback arises from the fact that only very small groups of subjects can be conveniently matched; and it has often been objected, on purely *a priori* grounds, that any small and heterogeneous group could probably be matched with fair success, whether the personality sketches were based on Rorschach responses, common observation, or indeed almost any relevant procedure. Thus, in my own cases subsequent enquiries from the judges often showed that they had been frequently guided, not by the sketch as a whole, but by some specific clue which had been inadvertently included by the writer and which was not really dependent on the test-results.

In the present research the total number of subjects was split into small groups, five to seven in each; and personality sketches were written from the orthodox Rorschach protocols. The results gave high figures: by Burt's formula the average coefficient was ·849; by Vernon's the combined figure was ·856. As a rough comparison, personality sketches based on the scoring by content were later read to judges of four of the groups. Here the coefficient was 1·000, i.e., every one of the subjects was matched correctly, although the sketches were decidedly short, and dealt with the cognitive traits quite as much as the orectic.

These figures, however, are not very trustworthy, partly owing to the difficulties in the matching technique already noted, and partly on account of the conditions of the experiment. For example, when writing up the sketches the investigator had already noted the contents of the responses. It was therefore difficult to be sure that the sketches, nominally based on a synthesis of the Rorschach category scores, had not also been influenced by attention to the content. Probably a more reliable way of approaching this problem will be to combine the Rorschach category scores so as to obtain weighted sums for the various traits that we are seeking to measure; and for this purpose we must have recourse to the more rigorous methods of factor analysis.

IV. A FACTOR ANALYSIS OF RORSCHACH CATEGORIES

Correlations between Categories. As Burt has pointed out, with the wide variety of Rorschach scores, two modes of factorial analysis are available—the first, based on the more familiar method of 'correlating traits,' the second on the comparatively novel method of 'correlating persons.' The latter, however, though highly suggestive, is not only less usual, but also more elaborate, and would require a paper to itself. Here, therefore, I shall confine myself chiefly to results obtained by the better known procedure.

Tetrachoric correlations have been calculated between the main Rorschach categories, thirty-six in all.[1] The two intelligence tests were also included in the series. If we accept the common assumption that the standard error of the tetrachoric coefficient is approximately 1·50 times the standard error of a product-moment correlation, then we may regard coefficients above 0·300 as statistically significant.

A number of the correlations had minus signs. As a result, on summing the columns to extract the first factor, it was found that eight of the categories produced negative totals, namely, Popular Responses (*P*%), Responses to Whole Blot (*W*%), Animal Responses (*A*%), Anatomy Responses (*AN*%), Responses to Form (*F*%), Beck's *Z*-score, the Modified *Z*-score, and, finally, the Average Time taken to make the first response. It would, of course, be quite impossible to start with a bipolar factor containing 29 positive and only 8 negative saturations. Accordingly, to secure positive saturations throughout the signs for these seven categories were reversed.

TABLE IV. FACTOR-SATURATIONS FOR RORSCHACH CATEGORIES AND INTELLIGENCE TESTS

Category	A. BIPOLAR FACTORS			B. GROUP FACTORS			
	I	II	III	*F*	*E*	*G*	*P*
Intelligence Test (Verbal)	·267	·516	—·152	·040	·590		
Intelligence Test (Non-Verbal) ..	·298	·675	—·085	—·011	·744		
Form Level Score (Klopfer)	·077	·805	—·189	—·260	·829		
Good Form (*F* + %)	·124	·731	—·170	—·240	·860		
Movement (*M*%)	·097	·597	·288	—·259	·800		
Few Anatomy (*AN*%)	·469	·510	·264	·235	·691		
Human Responses	·561	·362	·108	·378	·601		
Few Animal Responses (*A*%) ..	·595	·340	·143	·451	·546		
Architecture	·420	·112	·109	·417	·254		
Vocational	·475	·100	·103	·497	·219		
Form and Colour (*FC*)	·426	·088	·291	·384	·174		
Object	·619	·086	·050	·671	·245		
White Spaces	·546	·040	·008	·607	·159		
Few Form Responses (*F*%)	·163	·724	·288	—·248		·900	
Vista (three-dimensional)	·554	·364	·281	·400		·584	
Colour (*C*)	·142	·230	·623	·035		·469	
Surface Grey (*FY*)	·630	·172	·131	·525		·411	
Landscape	·748	·157	·288	·618		·521	
Abstract	·368	·113	—·415	·398		·336	
Clothing	·480	·088	—·004	·455		·178	
Few Popular Responses (*P*%) ..	·389	—·588	—·152	·495			·458
Human Details	·506	—·522	—·099	·513			·518
Proportion of Coloured to Uncoloured Responses	·166	—·306	—·109	—·077			·594
Low Average Response Time	·211	—·242	—·125	·327			·113
Clouds	·366	—·202	—·650	·446			·068
Animal Movement (*FM*)	·101	—·179	—·129	·007			·281
Colour and Colour Form (*C* + *CF*) ..	·344	—·125	—·560	·364			·077
Sex	·362	—·064	—·107	·342			·102
Cultural	·850	—·063	—·131	·780			·170
Low Modified *Z*-score	·135	—·998	·131	·128			·900
Low *Z*-score (Beck)	·429	—·702	·261	·372			·743
Bad Form (*F* —)	·377	—·650	·124	·545			·443
Few Responses to Whole Blot (*W*%)	·615	—·547	·510	·450			·694
Animal Details (*AD*)	·535	—·528	·036	·554			·514
Unusual Details (*Dd*)	·903	—·438	·075	·907			·473
Percentage of Large Details (*D*%) ..	·503	—·341	·344	·343			·505
Total Number of Responses	·962	—·266	·224	·897			·375
Plants	·482	—·046	·378	·400			·113

---

[1] The correlation table itself is too large to reproduce here. It will be found in the Appendix to my thesis (Table III, following p. 223).

1. *Bipolar Analysis.* The correlation table was first factorized by Burt's Method of Simple Summation, using successive approximation to secure appropriate communalities (16, pp. 461f.). Three factors, but no more, were found to be statistically significant, namely, one general factor, with positive saturations throughout, and two bipolar factors (cf. Table IV A). As has frequently been insisted (16), to be convincing any psychological interpretation of the statistical factors obtained in this way should, if possible, be supported both by internal evidence and by the evidence of external criteria. It will be convenient to examine the internal evidence first.

If, as is claimed, the Rorschach test covers the entire personality, and if further we accept the usual broad analysis of personality into two main aspects, cognitive and orectic respectively, then we might expect the two bipolar factors to represent *general intelligence* and *general emotionality* respectively, i.e., the one would contrast responses indicative of high intelligence with those indicative of low intelligence, and the other would contrast traits indicative of strong emotionality with those indicative of weak emotionality. As we shall see, owing apparently to the rather specialized diagnostic interests of those who have built up the method of scoring, this interpretation is not quite accurate. Further, the presence of a comprehensive *general factor* would seem to be somewhat unexpected to those who have hitherto offered a 'rationale' of the test.

(i) The first factor obtained with this procedure is a 'general' factor, having positive saturations ranging from 0·077 to 0·962. But, since so many traits have extremely low saturations, it contributes only 29 per cent. to the total variance.

From its very nature, a 'general' factor denotes efficiency or favourable performance in all the traits in the series from which it is obtained. A positive measurement for this factor must therefore represent favourable performance in the Rorschach test, as scored by the orthodox Rorschach method.[1] Can we give this factor any more precise interpretation in terms of familiar psychological characteristics? It cannot be identified with high general intelligence; for the items scored are by no means restricted to cognitive traits: they cover emotional and temperamental traits as well; and in some of the categories dullards may easily gain high scores. Moreover, the factor-saturations for the two intelligence tests are comparatively low, less than 0·300.

The highest saturations (over 0·8) are for Total Number of Responses, Number of Unusual Details, and Number of Cultural Responses. These figures, and a closer study of the other saturations, both high and low, suggest that the factor can probably best be interpreted as a general factor of *fluency of productive association.*[2]

This is, after all, not an unexpected result. Recent factor studies regard fluency of productive association as the essential factor entering into what, in traditional or popular language, is termed imagination (cf. 26, pp. 180f.); and it may be remembered that earlier workers used the inkblot test as a test of what (in terms of the old faculty psychology) was designated 'imagination.' It is so classified by Whipple; and Binet, Dearborn, Sharp, and Parsons all considered that its most marked characteristic was to display 'imaginative fertility.'

The category which gives the highest saturation of all is 'Total Number of Responses'; but, although the number of responses given by subjects with a high measurement for this factor is large, nevertheless (as the other saturations show) they do not consist of the easiest, most obvious, or most automatic types of response, such as are given by subjects possessing a ready reproductive association (e.g., responses relating to animals, anatomical resemblances, or mere 'form' responses). In the main they include a wide variety of imaginative, emotional, or cultural responses.

---

[1] The fact that some of the items have been reversed does not really invalidate this statement. Thus, as ordinarily scored, 'average time' denotes slow response; when reversed it denotes quick response. Similarly, what Beck scores as a large number of 'popular' or stereotyped responses, other workers would score in the reverse direction as a small number of 'original' responses, and so on.

It should be noted, however, that many scores are expressed as percentages of the total number of responses. Now we cannot expect all types of response to be equally available whether the total number of responses is small or large (for example, as Burt points out in dealing with associative responses, "the proportion of monosyllabic words to total number of words in the sample could hardly remain the same, when the sample is small and when it is large"). Hence the percentage-scores having low saturations must consist very largely of those characteristics that cannot increase at the same rate as the total number.

[2] It may perhaps be noted that other research students in the Department working with the Rorschach tests applied to English subjects, both adults and children, have also found a general factor of 'fluency of response,' e.g., A. Petrie and S. Cox.

Rorschach's account of the 'dilated' as contrasted with the 'coarted' type might seem in some degree to turn on the same type of mental process: but this is not altogether true. The usual measure of the 'dilated' type is a large score for 'the sum of colour responses *plus* the sum of movement responses' (cf. 20, p. 267); but these two categories do not, as a matter of fact, yield high saturations for this factor. Further, the 'coarted' or 'constricted' type is characterized, not so much by low fluency, as by "rigid intellectual control."

There seems to be a somewhat closer resemblance between this factor and the characteristics attributed by Jaensch to his 'eidetic' type. In his discussion of eidetic imagery he suggests that "certain selective principles operate on the higher levels of perception," and that what is involuntarily selected from the total number of stimuli presented and the associations they arouse "depends to a great extent on purely subjective factors such as meaning, values, conscious and unconscious wishes, interests, and previous experience." His two main types, the 'integrates' and 'disintegrates,' reflect the presence or absence of these selective tendencies. The differences in the products of their thinking are in fact due to the way they select. Thus the 'integrates' are the 'flexible thinkers'; they have 'rich associations,' and 'see meaningful wholes.' The 'disintegrates' are more objective; subjective valuation and experience have little effect on their perceptions; they are 'inflexible and analytic.' It would seem therefore that the subjects who gain high factor measurements for 'fluency' would be in many ways analogous to what Jaensch calls the 'integrate' type.[1]

(ii) The second factor contributes as much as 19 per cent. to the total variance. It is a bipolar factor; and the categories which have large positive or negative saturations are for the most part categories which relate to *cognitive* characteristics. Those obtaining high positive saturations are Klopfer's Form Level Score (which was designed to "get an objective scoring for the intellectual functioning" (19)), Percentage of Good Form, and Low Percentage of Form generally. Both the intelligence tests also yield high positive saturations. High negative saturations are given by Low Modified *Z*-score, Low Beck *Z*-score, Number of Bad Form Responses, Low Percentage of Responses to Whole Blot, and Number of Animal Responses. All the items involving 'details' (e.g., *D*%, *Dd*, *AD*, and 'Human Details') have fairly large negative saturations for this factor. It would appear therefore that this factor represents what Burt has called the '*generalizing*,' 'synthetic,' or 'integrative' type of attention[2] as contrasted with the '*particularizing*' or 'analytic' type.

The former type of mind is apt to be more abstract and intellectual, if not always more intelligent; the latter to concentrate rather on practical or isolated detail.[3] The contrast is in some respects analogous to what Rorschach, following Meumann, calls the *Wahrnehmungstypus* or *Erfassungstypus* (he defines it as 'mode of perceptual approach'); but the characteristics emphasized by the factor-saturations are by no means the same as those enumerated by Rorschach.

(iii) The third factor contributes only 7 per cent. to the total variance. The categories having high positive or negative saturations are to a large extent those commonly alleged to be indicative of the affective or *emotional* aspects of the subject's personality.

---

[1] Jaensch, E. R. (1930). *Eidetic Imagery*, pp. 92f.

[2] It should perhaps be pointed out that the word 'attention' is here used in the sense adopted by Stout and McDougall, namely, as the mode of consciousness essentially involved in acts of apperception (cf. Stout, G. F., *Manual of Psychology*, chapter on 'Attention,' esp. pp. 147f. of 3rd ed.). His account of the 'generalizing' and 'particularizing' types of attention is largely based on Meumann's distinction between *fixierende oder konzentrative Aufmerksamkeit* and *fluktuierende oder distributive Aufmerksamkeit* (*Experimentelle Pädagogik*, 1907, I, pp. 500f.). Earlier writers had distinguished three main forms of 'imagination': an 'abstracting' type, which seizes on isolated details; a 'determining' type, which fills out given details to make a determinate whole; and a 'combining' type, which creatively combines details not given in conjunction (Volkmann, *Lehrbuch der Psychologie*, I, pp. 470f.; Sully, *The Human Mind*, I, p. 364). Meumann argues that these are properties of apperception or, as he prefers to say, attention (*op. cit.*, I, pp. 241f.).

[3] This is also a contrast noted by Binet. It appears again in his account of his two replies in the tests of 'describing an object' and 'describing a picture' (1, 3: an accessible English account will be found in Myers' *Introduction to Experimental Psychology*, pp. 124f.). Myers himself, it may be added, later considered Binet's distinction between "the objective or practical type and the subjective or reflective type" of special importance in vocational guidance; Burt stresses its possible importance in educational guidance (selecting pupils for grammar schools as contrasted with technical or modern schools). Miss Sharp's main conclusion consisted in a confirmation of this twofold classification; and it reappears in Bartlett (7, p. 256).

Burt, in attempting to assess emotional tendencies from the Rorschach test results, follows the usual principles described by British psychologists[1] in discussing the influence of feeling on imagination and apperception; and lays stress chiefly on "the concrete contents of the interpretation." "These," he says, "often reveal the dominant emotional characteristics of the more active appercipient interests: in an apprehensive mood the child is more likely to see ghosts, skulls, or corpses; in a cheerful mood, to see fairies, toys, and Christmas trees. What may be called 'hedonic' and 'valuational' judgments are especially suggestive." Rorschach and his followers lay more stress on certain relatively abstract categories; their inferences are based on the allusions to 'colour,' 'movement,' 'shading,' 'white spaces,' and the like; and for the rest, they mainly rely on interpreting the ideas and images suggested as quasi-psychoanalytic symbols—the method so often stigmatized as the 'dream-book mode of interpretation.'

In Table IV A, it will be seen, the two largest positive saturations relate to the number of responses depending on Colour ($C\%$ and $C + CF$); on the other hand, the number of responses depending on Movement ($M\%$) yields a negative saturation, though the coefficient is not large.[2] So far this might suggest that the factor represents Rorschach's contrast between what he calls the 'extratensive' and the 'introversive' types. Bryn gives a tabular list of the characteristics supposed to distinguish these two types (23, p. 126); the chief characteristic of the 'extratensive' is said to be their 'unstable emotional life.' However, the highest positive saturation of all relates to 'clouds'; and there are fairly large positive saturations for items usually supposed to be characteristic of vague, introverted,[3] and autistic thinking (e.g., vistas, landscapes, abstract responses, etc.). This type of thinking is, of course, especially characteristic of the neurotic. And much the same conjunction of categories is reported by Guirdham, who found that a marked tendency to give Colour Responses often accompanies responses of the kind called 'Diffused Shading,' e.g., references to vague and relatively formless objects like 'clouds' or 'mist.'[4] Diethelm also noted a close relation between Colour Responses and Cloud Responses under the influence of adrenalin.[5]

All these considerations combine to suggest that this third factor tends chiefly to contrast subjects with *neurotic* and with *non-neurotic* tendencies respectively; and this inference is, as we shall see, confirmed by external evidence.

Correlations between Factor-measurements and Independent Ratings. With a view to obtaining a more objective check on the interpretation of the various factors, approximate factor-measurements were calculated for the bipolar analysis. The original scores were reduced to standard measure, and then weighted with the saturations for the three factors obtained by the bipolar method.

---

[1] Among British psychologists the writer who first laid stress on "the influence of emotional elements on imaginative combinations" was Bain: "vanity sets up pictures of admiring assemblies; anger, once roused, resuscitates objects in harmony with it; the emotion of terror gives a character to all the ideas formed under its influence: ghosts and hobgoblins fill the imagination of the superstitious" (*Senses and Intellect*, pp. 614f.); indeed, in *Mental Science*, p. 175, he proposes to "restrict the term 'imagination' to the productive process as differentiated by the presence of emotion." Sully similarly states "the presence of feeling gives a particular direction to the imaginative process: every feeling tends to reinstate those images that are congruous with it" (*op. cit.*, I, p. 378).

[2] Rorschach himself described the interpretation of 'movement responses' as "the thorniest problem in the entire experiment" (9, *Eng. Trans.*, p. 216). Earlier workers on apperception had contrasted 'colour,' not with 'movement,' but with 'form': Rorschach, chiefly it would seem for *a priori* reasons, proposed to substitute 'movement.' Burt, however, has given a good deal of evidence for maintaining that the change is not really justified (cf. 15, pp. 290f.); "the negative correlations are between interest in form and colour, not movement and colour." This is borne out by my own results. In the correlation table (27, appendix), it will be seen, there are small *positive* correlations between colour and movement, but comparatively high *negative* correlations between colour and form. Similarly in Table V most of the Form responses have negative saturations, and 'Few Form Responses' has a positive saturation. The highest correlation of $M\%$ is with the matrix test (+ 0·70); and its correlation with the group test of intelligence is also one of the largest (+ 0·40). This agrees with the view of many Rorschach workers that Movement Responses are indicative of intelligence.

[3] It should be noted that Rorschach's 'introversive' type is not (as many writers appear to suppose) altogether the same as the more familiar 'introverted' or 'inhibited' type: the latter in many ways corresponds more to his 'coarted' or 'constricted' type.

[4] Cf. Guirdham, A. (1936). 'The diagnosis of depression by the Rorschach test.' *Brit. J. Med. Psych.*, XVI, pp. 130–145, and refs.

[5] Diethelm, O. (1934). 'Personality Content in Relation to Graphology and the Rorschach Test.' *Proc. Ass. Res. Nerv. Ment. Dis.*, XIV, pp. 278–286.

This device for calculating factor-measurements is exceedingly convenient in practice. But it perhaps needs a word or two in justification. As a check on its validity, more accurate weights were also calculated by Ledermann's formula, namely: $FR^{-1} = (F'U^{-2}F + I)^{-1}F'U^{-2}$, where $R =$ the correlation matrix, $F =$ the matrix of factor-saturations, $U =$ the diagonal matrix of specific factor-saturations.[1] It was found that in their relative size they closely resembled that of the corresponding factor-saturations: in other words, the post-multiplication of F' by R^{-1} or its equivalent does not greatly affect the final result. The rank correlations between the two sets of figures are for the general factor ·82, for the first bipolar factor ·91, and for the second bipolar factor ·90. The correlations between the factor-saturations and the factor-weights are surprisingly high, and suggest that comparatively little loss of accuracy results from the substitution of saturations for factor-weights.

The approximate factor-measurements thus obtained were correlated (*a*) with the results of the intelligence tests and Cattell's test of fluency, (*b*) with the average of the judges' ratings for certain personality traits, and (*c*) with the personality ratings as estimated by Burt's procedure from the content of the Rorschach responses.[2] The correlations are shown in Table V. Only traits which provide at least one saturation over 0·30 are here included. It will be seen from the various figures that the conclusions suggested by the external criteria fully confirm those based on internal evidence.

TABLE V. CORRELATIONS BETWEEN FACTOR-MEASUREMENTS AND (1) TESTS, (2 A) JUDGES' ASSESSMENTS, (2 B) ASSESSMENTS BASED ON CONTENT OF RORSCHACH RESPONSES

TRAIT	FACTOR					
	I		II		III	
1. *Tests*						
Intelligence (Matrix) ..	−·132		·507		−·031	
Intelligence (No. 33) ..	·032		·456		·002	
Fluency (Cattell)	·485		·041		·194	
2. *Assessments*	A *Judges*	B *Content*	A *Judges*	B *Content*	A *Judges*	B *Content*
Intelligence	·196	·180	·502	·585	−·004	−·021
Perception of Relations ..	—	·096	—	·708	—	·082
Imagination	·530	·393	·146	·202	·092	·013
General Emotionality ..	·252	·371	·156	·186	·112	·188
Extraversion-Introversion ..	·149	·307	·169	·197	−·092	−·113
Neurotic Tendencies ..	·171	·242	−·212	−·319	·685	·702

For the first factor the highest coefficient furnished by the tests is the correlation with the test of fluency, and the highest coefficient furnished by the judges' ratings and the content ratings is the correlation with imagination: general emotionality and extraversion-introversion also yield fairly large figures. For the second factor the only large correlations are those with the two intelligence tests, the two ratings for intelligence, and the content rating for perception of relations. Nearly all the other correlations are statistically non-significant. The third factor yields a significant correlation with one trait only, namely, neurotic tendencies, as assessed both by the judges' rating and the assessment made by the investigator.

---

[1] Ledermann, W. (1939). 'A Shortened Method of Estimating Mental Factors by Regression.' *Psychometrika*, IV, pp. 109–116. Cf. Burt, C. (1937). 'Correlations between Persons.' *Brit. J. Psych.*, XXVIII, p. 85.

[2] For neuroticism I also attempted a twofold classification, based on my observations of the subjects during the course of the tests and interviews. The correlations derived for this classification are therefore biserial coefficients, and probably less trustworthy than the other figures.

2. *Group Factor Analysis.* A supplementary analysis was also carried out by Burt's Group Factor Method (16, pp. 477f.). The original matrix of correlations was divided into three submatrices in accordance with the lines of division suggested by the bipolar analysis; and a basic factor and three group factors were then extracted. The saturations are shown in Table IV B.

(i) The basic factor (F) places the categories in much the same order as did the general factor in the bipolar analysis, although the saturations have a wider range. The correlation between the two columns of saturations is 0·89. Hence it seems reasonable to regard the basic factor as representing fluency of association. It may perhaps be identified with Hargreaves' group factor of 'fluency' (derived from inkblot and other tests): he defines it as '*quantity* of imaginative production,' and considers that it confirms "the discovery of an Imagination factor by the authors of *A Study in Vocational Guidance*."[1]

It may be seen that a few of the categories have small negative saturations for this factor, namely, Good Form ($F + \%$), Klopfer's Form Level Score, Form Responses generally (F), Movement ($M\%$), Ratio of Coloured to Uncoloured Responses, and low Modified Z-score, as well as the Matrix test of intelligence. This probably implies that the large saturations assigned to the broad group factor (G), which covers these traits, really include saturations due to a narrower and more specialized supplementary factor, which, owing to the small size of the sample, cannot be significantly distinguished.

(ii) Of the three group factors the first (G) includes those traits that have positive saturations for the first bipolar factor and negative for the second. The persons belonging to this group tend naturally to generalize (they make free use of broad intellectual concepts), and are free from any marked neurotic tendencies. They thus show stable and well-organized personalities. The factor itself is perhaps best interpreted as a factor for integrated thinking, closely allied to intelligence.

Categories which are generally recognized as denoting intelligence (Klopfer's FL, $F + \%$, etc.), as well as Movement (which Hertz believes to indicate intelligence among children), have high saturations for this factor; and this interpretation seems confirmed by the saturations for the two intelligence tests. The fact that the test saturations are not higher still, however, appears to imply that, with the usual scoring-categories, it is scarcely possible to isolate from the Rorschach test a clear and reliable estimate of intelligence as ordinarily understood. As we have already seen, with a larger sample, the traits covered by this factor would probably further subdivide into two subgroups.[2]

(iii) The second of the group factors (E) includes categories, like Colour and Few Form Responses, which have positive saturations for both the first and the second bipolar factors. According to the orthodox interpretations, it should therefore denote a factor of emotionality. Except for 'Clothing,' which has a comparatively low saturation, the characteristics having high saturations in the group factor column—characteristics which include not only 'Colour' and 'Few Form' Responses, but Abstractions, Landscapes, Vistas, 'Surface Grey'—are highly suggestive of the type which Burt terms the 'unstable introvert,' a type whose emotionality is of the vague, dreamy, reflective kind.[3] Burt, for example, points out that, with picture tests, the 'unstable introvert' prefers "landscapes to portraits, scenes from dreamland rather than from reality" (15). These and other preferences suggest a close parallel to the kind of fanciful pictures which persons of this type tend to see in the inkblots.

---

[1] *Op. cit. sup.*, p. 24; cf. Burt *et al.*, *A Study in Vocational Guidance*, pp. 43–46, and 26, p. 181. Burt and other writers interpreted the underlying ability in terms of a positive endowment: this is implied by such terms as 'active imagination,' 'activity' of intelligence as distinguished from 'level' (Binet), 'cleverness' (Garnett), and the contrast drawn between 'imagination' and 'fancy' (Coleridge and Wordsworth, cited by Burt). Hargreaves, however, finds nothing corresponding to this positive conception, and believes fluency of association to depend essentially upon lack of inhibition: he compares it to the 'flight of ideas' in intoxicated and manic states. This, it may be remarked, comes near to Rorschach's account of 'dilation' as contrasted with 'coarctation.'

[2] It will be noted that the Rorschach test, as ordinarily scored, gives little or no information about special aptitudes. Some light is thrown on these by the scores derived from an analysis of contents. Miss Hillman, in an unpublished study with children, considered that verbal and practical ability, as well as productive and reproductive imagery (the latter including imagery and immediate and delayed memory), might all be tentatively assessed, but with no very high reliability, except perhaps in the case of productive imagination, "which the inkblot material was originally employed to test."

[3] Cf. Burt, C. (1917). 'The Unstable Child,' *Child Study*, X, p. 70 (Section on 'Association').

(iv) The last of the group factors (P) covers all those traits that showed negative saturations for the first bipolar factor. The subjects with high factor-measurements for this factor show a tendency towards 'particularization,' in the sense described above: their responses are determined, almost naïvely, by non-constructive mechanical association. This factor therefore seems to indicate a dominance of processes akin to what Burt calls 'reproductive association' and older writers 'passive imagination.'

The arrangement of positive and negative saturations in the second bipolar factor suggests that the traits entering into this group factor should also be subdivided. They appear to imply two distinguishable subtypes, namely, a more objective or practical type, who interpret quite impersonally the objective of concrete details which they see in the blots, and a mildly neurotic type, who interpret the details in a more personal or subjective way. Burt lays considerable stress on this latter tendency, and states that it "appears to occur much more frequently (*a*) in the feminine sex and (*b*) those persons of either sex who are of a psychoneurotic type and would often be described as suffering from 'inferiority' or 'superiority' complexes or even mild paranoid tendencies. Alike in tests of free association, in the interpretation of pictures, and in responses to ambiguous pictures and meaningless spot patterns, inkblots, and the like, they evince a marked tendency to give *individual* associations and interpret things in terms of their own *personal* experiences: more than half their replies are characterized by reminiscence or self-reference." The distinction appeared in Binet's experiments, and was most clearly marked in his tests of description. In his account of the results obtained from his two daughters, Marguerite (M) and Armande (A), "M," we are told, "writes much more frequently of her own past experiences than her sister: the difference is not so much due to memory as to interest. So, too, the words used by M refer more to her own person."[1]

However, with the present figures this distinction cannot be pressed. Within this set of traits the second bipolar factor shows only two large positive and two or three large negative saturations; the rest are decidedly small. It is therefore possible that these figures merely point to a certain degree of overlapping, or possibly to a couple of small special factors linking no more than two or three items in each case.

After the foregoing factors had been extracted, there still remained a few large residuals suggestive of a definite overlapping between certain categories. The more conspicuous suggest relatively specific correlations between 'Percentage of Colour Responses' and (*a*) 'Colour *plus* Colour-and-Form Responses' and (*b*) 'Few Whole Responses,' and again between 'Colour *plus* Colour Form Responses' and 'Modified Z-score.'[2]

The foregoing factor analyses help us to see far more clearly what it is precisely that we may to some extent infer from Rorschach scores. Although it is claimed that the Rorschach test covers the whole personality, it is obvious that it fails to give any clear information about many essential features. Indeed, it would seem that assessments based on it, particularly when scored in the usual way, are bound to be somewhat limited in value, first, because no single test and no single type of test-material can really do justice to all aspects of the personality; secondly, because the conventional scoring has evidently been appreciably biased by the special interests of those who have developed the test. Rorschach himself and most of the more enthusiastic champions of his test have been chiefly interested in the psychiatric field : hence their scoring has largely been concerned with the diagnosis of those forms of defective intelligence that are met with in the feeble-minded or the mentally diseased, and of the milder types of emotional instability that characterize the psychoneurotic and the psychopathic.

---

[1] Myers' summary, *Introduction to Experimental Psychology*, pp. 122–123. Some of the distinctions here drawn, it may be noted, are analogous to those observed by Bullough and others in their studies of artistic appreciation. Burt believes that, to a large extent, responses made in interpreting inkblots may be classified in much the same way as responses made in interpreting pictures, colours, and the like (cf. *How the Mind Works*, chapter on 'The Psychology of Art,' pp. 280f.). Miss Hillman, however, found that the correlations obtained, though usually positive, were seldom high: the largest correlations were obtained for the 'objective' and the 'associative' types. This seems to be in keeping with the inferences drawn above.

[2] In my thesis approximate saturations have been inserted to deal with the overlapping cases. With a larger group they could probably have been established more clearly by rotating the bipolar factors to fit the group factor pattern. Professor Burt, however, has suggested that an even more instructive form of rotation might be secured if the group factor pattern used as a guide was based upon the general structural pattern which has gradually emerged from the accumulated results of factor researches in the past. This, however, would seem to entail a larger number of independent assessments for the chief cognitive and orectic factors.

The factor analysis shows that *some* of the interpretations are undoubtedly justifiable, even if their reliability and validity are not very great. Thus the kinds of response said to be distinctive of high intelligence are, on the whole, accurately assigned. On the other hand, many of the assignments proved to be ill-founded or even fanciful. The suggestion that '*M*%' denotes 'imagination' is plainly negatived by its low saturation for factor I (0·097). The suggestion that *FY* denotes "lack of activity and listlessness" seems negatived by its high saturation for factor I. The statement that "Total *R*" is "the best indication of intelligence" seems contradicted by its negative saturation for factor II. The outstanding importance attached to *C*%, *M*%, and their sum and their ratio, is hardly warranted by the figures. And the reader who cares to compare the saturations for the several categories with their orthodox interpretations as given in the Rorschach guides will have little difficulty in discovering other instances where the usual diagnostic inferences can claim no measure of verification. Indeed, the large number of low saturations scattered throughout the table seems to suggest that many of the scoring-categories fail to confirm each other in the way that is claimed.

On the other hand, assessments based on analysis of contents seem far more promising. It is possible that a more eclectic type of scoring, based both on content and on the more suggestive of the formal characteristics, might, if subjected to a factorial study, render the test still more valuable for practical work. No doubt further studies on a much wider scale are required to substantiate the conclusions I have drawn. But the results so far obtained seem plainly to demonstrate the value of correlational studies along the lines here described.

Introspective Evidence. It is a sound principle that inferences as to the nature both of tests and of factors should be supported, not only by internal statistical evidence and by external observation, but also by introspective data, elicited from the subjects tested, and explaining the processes by which their overt responses were reached. Many Rorschach workers, after administering the test itself, conclude with a brief supplementary 'enquiry.' But this is usually directed to features in the blot, not to the mental processes of the subject. Its object is "to ascertain what part of the blot suggested the response, and whether it was determined by colour, shading, etc." : "questions should be asked only to clarify the scoring : enquiry into most responses is uninstructive and superfluous" (20, pp. 97-98). In research this restriction seems an unwise policy ; and, even in clinical work, where time is limited, it is not without serious risks.[1]

"Earlier investigations," it has been said, "on problems of this kind made the introspections the central feature of the experiment. In using indeterminate material (inkblots, spot-patterns, incomplete or ambiguous puzzle pictures, and the familiar alternating diagrams) the primary object of the investigators was to discover how, during the attention process, a relatively meaningless stimulus acquires meaning for the percipient. Trained most probably on Titchener's four-volume *Manual*, they would have learnt, from his very first experiment on attention (reporting on what is seen in puzzle pictures), that 'the appeal lies to introspection : . . . here, as elsewhere, the verdict of introspection must outweigh all preconceived opinions.'[2] Since then, however, owing largely to the reactionary criticisms of the behaviourist school, introspection has fallen into neglect. Too often the most characteristic responses are explained forthwith in terms of a small stock of alleged mechanisms, of which the percipient is supposed to be unconscious, when, in point of fact, he himself may be fully aware of the real reasons for his replies. Most frequently, of course, conscious and unconscious processes both play their part. But the former are nowadays too rarely investigated in attempts at validating and interpreting tests ” (10).

---

[1] Miss Hillman, for example, relates how a psychiatrist at a child guidance clinic, having applied the Rorschach test to a boy of thirteen, reported that "the bizarre responses (illustrated in detail) provided strong evidence for a diagnosis of latent schizophrenia" ; whereas a subsequent introspective study of the child's mental processes showed that "seven out of the eight 'bizarre responses' were words that the child had been writing out over and over again in a spelling-exercise given that morning in school." In other instances memories of quite recent incidents, especially at the cinema, "sufficiently explained a number of the children's more peculiar replies, which would otherwise have been quite inexplicable, and were dropped entirely when the test was repeated."

[2] Titchener, E. B. *Experimental Psychology : A Manual of Laboratory Practice*. I, Qualitative, Students' Manual, p. 109.

Many of the responses given by my own subjects could only be interpreted correctly by one who had already ascertained the particular features in their cultural background which so often prompted their replies[1] (cf. 27, pp. 33f., 93). Since the majority were untrained in introspection, the evidence available from this source was somewhat limited. Nor is a statistical journal the appropriate place for a detailed discussion of introspective studies. It may, however, be suggested that in future investigations the following points could usefully be borne in mind when investigating some of the questions raised by the correlational data and the factorial results.

As Titchener says, “ introspection knows nothing of faculties.” Hence, so long as we think of factors as abilities, we may gain little help. “ What introspection knows are processes.” Now, however meaningless the stimulus, the mental processes it arouses will inevitably be affected by “ the tendency of everything psychical to secure meaning.”[2] And “ apperception,” as Stout reminds us, “ is the process whereby a presentation acquires significance for thought,” or, in a word, “ secures meaning.”[3] Introspection thus seems essential if we are to obtain a first-hand knowledge of the various forms this process may take. With the blots, for example, some subjects will explain how the “ meaning ” nearly always presents itself immediately, after the briefest glance ; others report that it only comes after an appreciable interval spent on what Bartlett would call “ rummaging about,” “ searching among ancient memories now half forgotten.” We have therefore to distinguish between factors operating in and during the apperceptive process itself, and those that operate outside it or before it ; and between those that are conscious, and take time to evolve, and those that are more or less unconscious, and may consequently operate almost at once. Whether clearly conscious or not, all such processes will have a threefold aspect—cognitive, affective, and conative. The cognitive report, for example, may show that the subject in question nearly always finds the *name* comes first, and he then develops the implications of the name ; other subjects will say that they first picture some related incident or object, and experience considerable difficulty in finding suitable words or names. But what suggests the names, the incidents, or the objects ? On the affective side, an accompanying hedonic tone seems always to be present, though it may not be mentioned until an introspection is requested. If, as we are told, attention to dark shading commonly denotes an unpleasantly toned ‘ anxiety,’ and attention to light shading a pleasantly toned ‘ euphoria,’ introspection should, with trained psychological students at least, be able to confirm it. On the conative side, lack of fluency may or may not be due to inhibitions, or to the fact that the subject has to make what Bartlett calls an “ effort after meaning.” Free association and other analytic techniques may help to disentangle the deeper and less conscious influences. It may, I think, be fairly predicted that even the simplest experiments along these lines will often throw more light on the factors actually at work than either speculative interpretations or a mass of unexplained statistical data. Only, therefore, when we fully understand the *general* psychology of the Rorschach test shall we be able to use it scientifically for individual assessments (10).

[1] This, I venture to suggest, would seem to imply that, in child guidance work, it is highly inadvisable to accept interpretations of children’s replies, put forward by psychologists or psychiatrists who are often quite unfamiliar with the daily life of the children in question and may themselves have even grown up in another country. It certainly confirms the criticisms so often made by teachers on reading reports by psychologists of German origin on English schoolchildren brought up in a London slum.

[2] Ewald, W. *Prinzipien der Denkpsychologie*, p. 619.

[3] *Analytic Psychology*, II, p. 110.

V. SUMMARY AND CONCLUSIONS

1. A hundred Indian students were given the individual Rorschach test, two intelligence tests, and Cattell's test of fluency. The Rorschach test was scored, first, by Beck's standard method, and, secondly, by Burt's method of rating the contents of the responses in accordance with a schedule of personality traits, such as is commonly used in individual psychology. Ratings for the various traits were also obtained from independent judges, who were close acquaintances of the subjects tested; and these were used as validating criteria.

2. Correlations were calculated between the judges' assessments and (*a*) the individual Rorschach categories (classified according to Beck's scheme), and (*b*) the ratings based on an analysis of content. With few exceptions, the Rorschach categories, as scored by the orthodox procedure, gave figures that were statistically non-significant. On the other hand, the method of rating by content furnished consistently high coefficients.

3. Correlations between the 36 Rorschach categories were factorized by (*a*) simple summation and (*b*) the group factor method. The interpretation of the factors obtained was based partly on the internal evidence of the saturations themselves, partly on the external evidence supplied by independent assessments, and partly on introspective evidence obtained during the supplementary enquiry. The bipolar analysis indicated, first, a general factor of fluency, and, secondly, two bipolar factors concerned with cognitive and with emotional traits respectively. The first of the two bipolar factors appeared to contrast responses of a synthetic or generalizing type with those of an analytic or particularizing type; the second appeared to contrast neurotic with non-neurotic tendencies. The group factor analysis suggested a basic factor of fluency, and three group factors which might be interpreted as representing very approximately general intelligence (an integrative or generalizing factor), general emotionality (with a marked bias towards introversion), and a relatively mechanical tendency towards the associative reproduction of particulars.

4. To provide a check on the inferences drawn from the factor-saturations, approximate factor-measurements were calculated for each factor; and these were correlated with (*a*) the results of independent tests, (*b*) judges' ratings for the more relevant traits, and (*c*) assessments based on an analysis of the content of the responses. On the whole the correlations appeared to corroborate the tentative identifications of the several factors.

5. The results of the factor analysis confirmed several of the orthodox interpretations regularly suggested for some of the commoner types of response, particularly for those alleged to be indicative of intelligence and of neurotic tendencies. Many of the more formal categories, however, appeared to have little or no general psychological significance; and the inferences commonly drawn from responses referring to colour, movement, and the like seem to rest more on preconceived hypotheses than on verifiable evidence. It is concluded that an analysis of content, based on generally recognized psychological principles, and expressed in terms of generally recognized psychological factors, would be far more valuable for practical purposes.

REFERENCES

1. Binet, A., and Henri, V. (1895). 'La psychologie individuelle.' *L'Année Psychologique*, II, 411–465.
2. Dearborn, G. (1897). 'Blots of ink in experimental psychology.' *Psych. Rev.*, IV, 390–391.
3. Binet, A. (1903). *L'étude experimentale de l'intelligence*. Paris: Alcan.
4. Whipple, G. M. (1910). *Manual of Mental and Physical Tests*. Test 45. 430–435.
5. Burt, C., and Moore, R. C. (1912). 'Mental differences between the sexes.' *J. Exp. Pedag.*, I, 273–284, 355–388.

6. Pyle, W. H. (1913). *The Examination of School Children.* New York : Macmillan.
7. Bartlett, F. C. (1916). ' An experimental study of some problems of perceiving and imagining.' *Brit. J. Psych.*, VIII, 222–267.
8. Parsons, C. J. (1917). ' Children's interpretations of ink-blots.' *Brit. J. Psych.*, IX, 74–92.
9. Rorschach, H. (1921). *Psychodiagnostik : Methodik und Ergebnisse eines Wahrnehmungs-diagnostischen Experiment.* Bern : E. Bircher.
10. Burt, C. (1934). *The Rorschach Test.* Laboratory Notes, University College, London. (Revised 1945.)
11. Kerr, M. (1934). ' The Rorschach test applied to children.' *Brit. J. Psych.*, XXVII, 51–59.
12. Vernon, P. E. (1935). ' On the significance of the Rorschach test.' *Brit. J. Med. Psych.*, XV, 199–207.
13. Vernon, P. E. (1935). ' Recent work on the Rorschach test.' *J. Ment. Sci.*, LXXXI, 1–27.
14. Vernon, P. E. (1936). ' The matching method applied to investigations of personality.' *Psychol. Bull.*, XXXIII, 149–177.
15. Burt, C. (1939). ' The factor analysis of emotional traits.' *Character and Personality*, VII, 238–254, 285–299.
16. Burt, C. (1940). *The Factors of the Mind.* London : University of London Press.
17. Burt, C. (1945). ' The assessment of personality.' *Brit. J. Educ. Psych.*, XV, 107–121.
18. Beck, B. J. (1945). *Rorschach's Test.* New York : Grune and Stratton.
19. Klopfer, B., and Kelley, D. (1946). *The Rorschach Technique.* New York : World Book Company.
20. Rapaport, D. (1946). *Diagnostic Personality Testing.* Chicago : World Book Company.
21. Cattell, R. B. (1946). *The Description and Measurement of Personality.* New York : World Book Company.
22. Eysenck, H. J. (1947). *Dimensions of Personality.* London : Kegan Paul.
23. Bell, J. E. (1948). *Projective Techniques.* New York : Longmans, Green and Co.
24. Ramzy, I., and Pickard, P. (1949). ' The reliability of Rorschach scoring.' *J. Gen. Psychol.*, XL, 3–10.
25. Goodenough, F. L. (1949). *Mental Testing.* New York : Rinchart.
26. Burt, C. (1949). ' The structure of the mind : a review of the results of factor analysis.' *Brit. J. Psych.*, XIX, 176–199.
27. Sen, A. (1949). *A study of the Rorschach test.* Doctorate Thesis, University of London Library.

GROUP FACTOR ANALYSIS

By CYRIL BURT
University College, London

I. *General Problem.* II. *Non-Overlapping Group Factors :* A. *Three Group Factors.* B. *More than Three Group Factors.* C. *Two Group Factors.* D. *One Group Factor.* III. *Overlapping Group Factors.* IV. *Summary and Conclusions.*

I. GENERAL PROBLEM

Aim. In the following paper my main purpose will be to state the essential principles underlying what I have called the ' group factor method ' of factor analysis, and to describe the simplest working procedures available for the commoner types of problem that occur in actual practice. This will give me an opportunity of replying more fully to criticisms of the method raised by earlier writers, and of answering relevant questions received from recent correspondents. At the same time I shall endeavour to explain in what ways the underlying concepts resemble, and differ from, those involved in the related notions of ' simple structure ' and ' bifactor analysis.' The fact that the recent edition of Professor Thomson's book now includes an ' Addendum ' on the ' bifactor method ' has aroused a special interest among British workers ; and has led to several enquiries about its relations to previous procedures, which I shall attempt to answer so far as my own limited experience of it permits.[1]

Need for a Group Factor Method : (1) With Physical Measurements. In several previous contributions I have described how the notion of factorizing a correlation table into group factors originally arose (7, 8, 9, and refs.). Here I need only recapitulate the history of the concept quite briefly.

The need for some such procedure was encountered first in connexion with the analysis of physical measurements. To make the problem clear, let us begin with an imaginary instance. Suppose we have taken four representative measurements for the size of the human body (e.g., height, weight, chest breadth, and arm length) and three for the size of the head (e.g., length, breadth, and height) ; suppose, further, that we have found large positive inter-correlations between the first four traits, and large positive inter-correlations between the last three traits, but zero or non-significant cross-correlations between the first four and the last three : in such a case we should, I imagine, conclude that there were two distinct and separate growth-tendencies at work—one for the growth of the body and another for the growth of the head.

So obvious a conclusion would be drawn regardless of any thought or theory of factorial analysis : it would have been unhesitatingly inferred by Galton, Pearson, or any of their followers half a century ago. It is, however, convenient to have a more general name for such hypothetical tendencies ; we may therefore agree to call them 'factors.' And, since they are factors that enter into limited groups of traits only, it is natural to refer to them as 'group factors.'

---

[1] *Factorial Analysis of Human Ability* (1948), pp. 341f. In this country the bifactor method seems almost unknown. The only large-scale enquiry in which it has (to my knowledge) been used is an unpublished research on ' Temperamental Factors,' recently carried out by Mr. C. J. Adcock (Ph.D. Thesis, University of London, 1937). Thurstone, in his recent work (*Multiple Factor Analysis* (1947), p. viii), complains that Holzinger and Harman " dispose of the [his] simple-structure concept in a footnote " (10, p. 102) ; but he himself dismisses the bifactor method with only a parenthetical sentence in his preface. In my own view, as I have said elsewhere (7, 9), the different methods put forward by different writers should not be regarded as rival competitors : each may have its special merit for the purpose for which it was contrived. Hence I hope it will be understood that my criticisms relate primarily to the particular kind of problem on which my fellow-workers and I have been principally engaged.

If, as psychologists, we are interested more particularly in the growth of the head or skull, we might go on to enquire which of the three head measurements, or better still what weighted sums of all three measurements, would give the best indication of skull-growth, and indirectly, we might suppose, of brain-growth. By means of such weights we could obtain explicit 'factor-measurements' for our 'group factor.'

Now, when we turn to an actual table of such correlations, for example, the table for physical measurements discussed by Pearson and Macdonell in 1902,[1] the pattern discovered is not so clear cut : we find that, while the inter-correlations within the two main groups are both positive and large (averaging 0·67), the cross-correlations are by no means negligible ; though relatively small (averaging only 0·27), they are all positive and all significant. Thus, as Pearson expresses it, *both* 'like' *and* 'unlike' traits are positively correlated, although the correlations between 'like' traits are much higher. Evidently two independent and non-overlapping group factors will not suffice to explain the figures actually observed.

Pearson himself, it will be remembered, suggested extracting as many uncorrelated 'factors' (or 'index characters,' to use his own term) as there are correlated traits, and choosing for that purpose the 'principal axes' of the correlation ellipsoid. In that case all factors except the first would necessarily be bipolar. However, on examining that portion of his table which gave the cross-correlations, it was observed that, curiously enough, the figures were almost exactly 'synclinal' or 'hierarchical,' i.e., they virtually formed an oblong *submatrix of rank one* (see 12, Table I, p. 104). It therefore seemed uneconomical to use seven factors to explain these particular figures when one factor alone would serve. And in fact a very close fit to the whole table was obtained with three factors only, one 'general' and two 'group' (*ibid*., p. 118).

(2) *With Mental Tests.* Similar correlation patterns were from time to time observed in the early results obtained for mental traits. But it was not until about 1924 that the idea of group factors became widely accepted by psychologists.[2]

In my first paper on intelligence tests (1909) I found evidence for "groups of allied tests to correlate together" : thus, after deducting the effects of general intelligence, there still remained significant or nearly significant residuals for the inter-correlations (*a*) between tests of sensory discrimination, (*b*) between tests of motor co-ordination, (*c*) between sensori-motor speed tests, and (*d*) between tests and assessments involving memory (1, pp. 144, 164, and Tables V and VI). These results were confirmed and extended in an investigation carried out with Mr. Moore on much larger groups of schoolchildren, tested by means of group tests, which included 'higher mental processes.' Independent corroboration seemed to be afforded by Simpson's work with adults.[3] Simpson applied 14 tests, representing different cognitive levels (similar to the five 'levels of mental process' tested in my own investigation), to 37 adult males ; he also found, in addition to 'general mental ability,' 'certain relatively specialized capacities' belonging to particular categories or groups (very similar to those I had already reported), namely, (*a*) sensory discrimination, (*b*) motor control. (*c*) memory and association, (*d*) quickness and accuracy of perception, and (*e*) selective thinking. Simpson, however, based his inferences merely on a comparison of the observed coefficients, and made no attempt to eliminate the general factor or test the residuals for significance.

Our conclusions were at first strongly criticized by Spearman and others. Not without justice, he objected, first, that the statistical techniques employed were crude and inconclusive, and secondly that the numbers tested were too small for the results to be accepted as fully significant. Judged by the intercolumnar criterion, he found that my own data gave a mean corrected correlation of over 1·00 in both investigations, and Simpson's one of 0·96. Moreover, in common with other

[1] *Biometrika*, I, 1901–02, pp. 177–207. In this enquiry the dimensions of body and head actually measured were not quite the same as in the hypothetical example I have taken in the text, since the investigators had to choose their figures from the Bertillon scheme of measurements collected by Scotland Yard from criminals.

[2] See *Report on Psychological Tests of Educable Capacity*, Board of Education, 1924. In Appendix IX, the majority of psychologists agreed in accepting, at least provisionally, a hypothesis including both a general factor and a number of group factors. A review of the chief group factors that now seem to be fairly well established will be found in *Brit. J. Educ. Psych*., XIX, 1949, pp. 100–111, 176–199.

[3] 'Correlations of Mental Abilities,' *Columbia University Contributions to Education*, No. 53, 1912

writers at that date, he doubted whether the distinction between 'like' and 'unlike' traits, so conspicuous in the case of physical measurements, could be expected to hold good in a field so different as that of mental testing, "unless the 'similar' tests were so similar as to be virtually identical."

American criticisms were less emphatic. Freeman, for example, observed that "Simpson presents his results in the form of a table similar to that which Burt reported, and the order of the tests bears out the same conclusion which was drawn from Burt's": nevertheless, he had "doubts as to the validity of the classification" ("according to mental categories"), largely because "the correlation between traits in different classes is often much closer than the correlation between traits in the same class. . . . If Spearman's two-factor theory is correct, special capacities do not fall into groups."[1]

(3) *With Educational Abilities.* Conclusive evidence for the existence of such 'group factors' was first obtained in assessments of educational abilities. Here written tests could be employed; and data could consequently be secured from samples large enough to render the first residuals fully significant and the second residuals definitely non-significant—at least in the earlier studies.

To estimate the correlation of each trait or test with the hypothetical factors, a method of 'simple summation' was employed, based originally on the assumption that the 'observed average' of a given set of correlations could be identified with the 'true average.' In dealing with residual correlations left after the effects of the first or general factor had been eliminated, two courses seemed open: "we may look either (*a*) for limited group factors, shared only by similar sets of tests, whose influence will be solely positive (the natural assumption in analysing intellectual abilities), or (*b*) for supplementary general factors, ambivalent in their nature, which will account not only for specific resemblances but also for specific antagonisms (an assumption that appears perhaps more plausible in analysing emotional and temperamental differences). Moreover, these alternative aims will, as a matter of fact, affect our conception of the general factor, and the way in which we estimate it. . . . When we come to the secondary factors, with assumption (*a*) the obvious working-procedure will be to take the significant positive correlations" (i.e., the clusters of positive residuals) "in subgroups, and factorize these just as we should complete tables."[2]

It was by this latter procedure ('Method *a*') that, working in collaboration first with Mr. R. C. Moore and later with Miss M. Bickersteth (3), I obtained what seemed to be unquestionable evidence for the presence of special cognitive abilities in both mental and scholastic tests—abilities of a kind which at that time Spearman and others strongly denied on the ground that they resembled the old-fashioned 'faculties.' The first explicit confirmation of our results came from one of Spearman's own pupils, Miss N. Carey, who based her deductions on school marks. "With actual tests of school abilities," she writes, "Mr. Burt has found[3] a 'general educational ability,' combined with 'specific educational abilities' (arithmetical, linguistic, literary, and manual), very similar to the combination of the 'hypothetical general factor' with subordinate 'group factors' that he found in his intelligence tests"; and she points out that her correlations for school marks indicate much the same combination of three types of factor, namely, in addition to (i) the 'general' and (ii) the 'specific factors,' (iii) a series of 'group factors' : (in her own data, these last comprised a verbal factor for verbal subjects, a motor factor for handwork subjects, and probably a third factor for informational subjects). To simplify the computations, however, she proposed to modify my procedure by assuming that the residuals on which any one group factor is based are approximately all the same in size. Incidentally she observes that her results are in marked conflict with the assumptions made by Spearman and Hart in examining my earlier paper, namely, that the high intercolumnar correlations obtained from tables like my own commonly demonstrate that there is *no* such clustering of high coefficients as the presence of group factors would imply. She herself actually obtains *negative* correlations between columns when these relate to tests that contain different group factors and low *g*-saturations ; and she uses this result to justify the way she has partitioned her table.[1]

---

[1] Spearman's most detailed discussion of Simpson's result is to be found in his *Abilities of Man* (6, pp. 145f.). He there applied his later 'tetrad-difference criterion' to Simpson's correlation table (it was the only large table for which this criterion was worked out in detail); and concluded that the resulting graph "displays one of the most striking agreements between observation and theory" [the theory of a single common factor] "ever recorded in psychology": in a footnote, however, he admitted that there are "some very few cases . . . more than five times the probable error" (and a still larger number, it might be added, more than *three* times the probable error). Freeman's criticisms, embodying earlier comments, will be found in his *Mental Tests* (1926), pp. 77–80.

[2] *Annual Report of the Psychologist to the L.C.C.* (1914–15), Appendix I: cf. 7, p. 306 and pp. 62–68 below. Here I should perhaps explain that the phrases used to designate the two alternative procedures have varied considerably. The full titles should be 'summational analysis into general and bipolar factors' and 'summational analysis into general (or basic) and group factors': but these circumlocutions are exceedingly cumbrous. In early memoranda on working methods (quoted in 7, p. 306, and 8, p. 356) the two alternatives were simply described as 'Method *a*' and 'Method *b*' respectively. Later, 'Method *a*' was called 'analysis by submatrices.' Nowadays the procedures are, as a rule, referred to as 'bipolar' and 'group factor' analyses, it being tacitly assumed by most British factorists that there will usually be a 'general' or 'basic' factor as well. The use of different designations in the numerous published and unpublished researches, as well as the need for slightly different working formulæ in different cases, has, not unnaturally, prevented non-statistical readers from noting that the essential principles have remained the same.

[3] Miss Carey is referring to results reported in the *Annual Report of the Psychologist to the L.C.C.* (1914) (briefly summarized in 3, pp. 34–37). Appendix I of the *Report* contained a short account of the derivation of the formulæ (afterwards included in *Mental and Scholastic Tests*, 2nd ed., pp. 271–275).

In a number of later researches and theses, the group factor method was regularly employed. But for the most part the working procedures adopted were improvised to meet the special needs of each particular enquiry ; and, as the interest lay rather in the psychological results obtained than in the mathematical devices used, little trouble was taken at the time to establish them systematically on a comprehensive basis. Indeed, during the 1920s, it would have been commercially impossible in this country to publish a book containing numerous correlational formulæ and expensive tables except with the assistance of the London County Council ; and the Council, quite properly, was more concerned to publish accounts of the practical conclusions reached than discussions of statistical procedures. As a result, during and since the war, younger psychologists have tended to substitute the methods described in the more accessible American publications, particularly the 'centroid method' and the device of 'simple structure.' Nevertheless, it has apparently proved difficult to dispense altogether with the 'general' or 'basic' factor, as 'simple structure' seems to require.[2] And, partly for this reason no doubt, many requests have been recently received for a factorial method which would take such a factor into account. However, before attempting to illustrate in detail the working procedures suggested, it will be advisable to formulate a little more precisely the underlying hypothesis and the fundamental principles on which all of them are based.

The Hypothesis Involved. The early work I have just described seemed to make it more and more evident that, in all researches on individual psychology and especially in the field of cognitive and scholastic testing, a possible interpretation to be kept constantly in mind was what I have termed the 'group factor hypothesis.' The phrase is convenient, but curtailed. I use it to denote the hypothesis that the traits tested or measured can best be analysed into (i) a basic or general factor, entering into *all* the tests, and alone accounting for the cross-correlations between the several groups, and (ii) a set of group factors, either overlapping or non-overlapping, each

[1] Carey, N. (1916). *Brit. J. Psych.*, VIII, p. 180. Miss Carey's procedure departs in one or two minor respects from my own. In previous discussions both Pearson and Spearman had criticized the way I had proposed to deduce, from restricted parts of the table only, a general factor which by definition was to cover the whole table.

Pearson insisted (and I think rightly, so far as a rigorous proof is concerned) that the first step should be to show that significant residuals were left, even when the closest possible fit to the table as a whole had been obtained (i.e., to use factorial terminology, that the general factor should be calculated from the entire table by the method of least squares, or some equivalent procedure). Further, he argued (as mathematical critics of current factorial methods are still inclined to do) that the leading diagonal should contain unity, not reduced self-correlations. When we try to discover the minimum self-correlations from unrepeated tests, we run the risk of trying to deduce more independent variables than our table permits, and we further render any valid test of significance doubly difficult.

Spearman insisted that the general factor ought rather to be based on some external criterion, e.g., teachers' assessments or selected tests used as reference values ; and he considered that a true 'hierarchy' should show *equal* decrements, not rows of coefficients diminishing in proportion. Accordingly, since Miss Carey was venturing to criticize Spearman's own conclusions, it was necessary for her to adopt his general procedure. Thus, in calculating the saturation coefficients (" the correlation of each school subject with *g* "), she used Spearman's method, which derived the 'saturation' not from the table of observed inter-correlations, but from two special tests assumed to give pure measures of *g*. In her main argument, too, she assumed that the hierarchy would be additive, though towards the end of her paper she is inclined to adopt my own procedure (cf. pp. 175 and 182).

[2] Vernon, for example, observes that, " in spite of the fact that no Navy or Army psychologist was an ardent disciple of Spearman, . . . over and over again our test inter-correlations have yielded a general factor running through all the tests and covering 30 to 40 per cent. of the variance, together with two, three, or more group factors " ('Statistical Methods in the Selection of Navy and Army Personnel,' *J. Roy. Stat. Soc.*, VIII, 1946, p. 145).

of which is primarily limited to a specific cluster or *group* of tests (presumably measuring 'like' or 'similar' mental functions), and in all of which the significant saturations are essentially positive.

In this hypothesis there was nothing new except the attempt to define the relevant concepts with sufficient precision to permit of rigorous statistical verification. Some such scheme of factors is already implied in the traditional attempt to classify mental faculties or functions into comprehensive *genera* and subordinate *species*. It is still more clearly implied in Galton's doctrine that an individual's achievements in a particular sphere may be explained in terms of some 'specific aptitude' for that sphere of work, superposed upon a 'general ability' required in every sphere.

A Working Procedure: General Principles. Having defined our terms, our next task is to deduce the verifiable corollaries of this particular hypothesis, and see how, in any concrete case, its consequences can be tested by means of the data actually obtained. Where more than one factor was concerned in any given test, it was assumed that, whether cognitive or conative, 'general,' 'basic,' 'bipolar,' or 'group,' the factors would combine (like velocities and forces in ordinary dynamics) in accordance with the 'parallelogram law.'[1] From this it follows that, if the factors postulated are independent and uncorrelated, the correlation between any two tests, r_{xy} say, could be expressed as a product-sum of the factor-weights or 'saturations,' i.e.,

$$r_{xy} = r_{xb}r_{yb} + r_{xc}r_{yc} + \dots,$$

where the hypothetical correlations between the tests and the factors (r_{xb}, r_{yb}, etc.) are taken as weights (see Appendix, p. 72, equation iv). This in turn implies that any observed correlation matrix is regarded as analysable into a sum of superposed 'hierarchies' (i.e., matrices of rank one), each of which may cover either *all* the traits tested or only *some* of the traits tested (8, p. 344).

As applied to the group factor hypothesis, these assumptions lead to two convenient requirements which any working procedure should fulfil. (i) First, "so far as possible, all the *significant* residuals (on which the supplementary factors were to be based) should be positive, never negative." (ii) Secondly, "all the *non-significant* residuals should for any given test (i.e., for any given row or column of correlations where only one factor was involved) sum to zero."

Each of these conditions requires a comment.

(i) To most teachers and educationists, it would certainly appear clumsy and far-fetched to classify cognitive traits in bipolar terms, and to postulate 'abilities,' or at any rate 'factors,' which have about as many negative saturations as positive. Indeed, with a few notable exceptions, both logicians and scientists have commonly been averse from dichotomous classifications when others are available.[2] Accordingly, in researches like those I have described, it seemed simpler and more intelligible to postulate positive group factors for verbal, arithmetical, and manual tests respectively (in accordance with 'Method *a*'), instead of classifying the various performances first into 'verbal' or 'non-verbal,' and then subdividing the 'non-verbal' (and sometimes quite illogically the 'verbal') into 'manual' and 'non-manual,' and so on (as required by 'Method *b*'). No doubt, even with educational abilities the occurrence of an occasional residual with a negative sign is not of necessity inexplicable. But it should be regarded as exceptional.

(ii) If we desire to judge 'goodness of fit' by the principle of least squares, then the closest fit will be obtained by taking a *weighted* sum of the non-significant residuals: for then our aim will be to minimize the square-sum of all the non-significant discrepancies. This entails a rather elaborate process of successive approximation. In practice, however, it appeared sufficient for most purposes to substitute the principle of 'simple (i.e., unweighted) summation,' similar to that adopted in calculating an ordinary average or a mean deviation. This yields a direct and simple routine procedure for obtaining factors of the type required: (see Appendix, pp. 72–73 below).

[1] This assumption was first put forward in attempting to analyse the 'general factors and specific factors underlying the primary emotions' (cf. *Brit. Ass. Ann. Rep.*, 1915, pp. 694–696): it was an obvious corollary to McDougall's view of emotions as essentially conative 'forces.' But almost immediately it became plain that, from a formal standpoint at any rate, cognitive capacities could be treated in precisely the same way.

[2] Jevons, F. B. (1900). *Principles of Science*, ch. XXX; Joseph, H. W. B. (1931). *Introduction to Logic*, pp. 121–122.

Accordingly, let us now consider what are the most convenient procedures for factorizing a given correlation table along these lines, and how we can best test the 'goodness of fit' furnished by factors obeying the scheme described. We may take the simplest form of the hypothesis first of all, namely, that which assumes that there is no overlap between the various group factors. As we shall see, the double principle just described leads to slightly different formulæ and slightly different working methods according to the number of group factors required.

II. NON-OVERLAPPING GROUP FACTORS

Partitioning the Correlation Table. The device on which the proposed procedure depends is that of 'condensation.' After the observed table of correlations has been appropriately rearranged and partitioned, in accordance with the grouping of the tests, the various figures or 'elements' in each submatrix are summed, so as to reduce the entire matrix of observed correlations to one of lower order. Provided the groups show no overlapping, this can then be treated as an approximate matrix of rank one, except for those sums which now appear in the diagonal.[1] But, before we can proceed to the actual calculation of the group factor saturations, we must first decide how the original table is to be partitioned, in other words, how the tests or traits are to be grouped.

Thomson, in an appendix on the bifactor method, added to the last edition of his book, mentions three alternative suggestions. But one or two others may also be put forward.

(1) *Subjective Classification.*—Probably the first and most natural suggestion would be that the grouping should be made "subjectively, by considering the nature of each test, and putting together the memory tests, or tests involving number, and so on" (11, p. 341). As numerous theses show, this is a favourite method with the youthful investigator. Yet it seems highly inadvisable. It provides too easy a way of appearing to establish the very hypothesis the investigator hopes to prove. The purpose of factorization is to allow the analysis to reveal, quite automatically and objectively, what are the actual groupings. If, however, we start by *assuming* (say) that all our memory tests form one group and all our number tests another, then we shall be basing the method of calculation on the very conclusion which the method of calculation ought automatically to reveal.[2]

(2) *Coefficient of Belonging.* Holzinger prefers to calculate a 'coefficient of belonging.' This is a somewhat slow and laborious preliminary; and, as ordinarily used, it does not always lead to a unique subdivision. Most investigators seldom calculate every possible value for the crucial coefficient; and it is then all too easy to stop when the results correspond with the investigator's preconceived desires.[3]

(3) *Correlation Profiles.* A third method is to make a graph or profile of each row of correlations, grouping those tests with similar profiles. This procedure has been most strongly advocated by Tryon. Thomson states that he himself prefers to compare, not the contour of the profile as a whole, but "only the peaks of each row." The device undoubtedly throws light on the structure of the correlation table[4]; but, in many cases if not in most, it does not yield a very satisfactory

[1] Numerous short-cuts depend upon this general principle, e.g., the 'sum method' of bipolar analysis (see, for example, this *Journal*, II, i, p. 61).

[2] Of course, it is possible that the whole research may have been planned to test some particular hypothesis, involving a specified grouping of the traits. But in that case it is essential to show that the group factor matrix reached on this hypothesis is the only hypothesis that gives a satisfactory fit, or at least gives a better fit than any other hypothesis. In the simpler type of problem this can be done quite simply by showing that a bipolar analysis leads automatically to the grouping assumed.

[3] This has been the experience of a good many workers in this country, as will be seen from their researches. One of the most recent and convincing set of illustrations will be found in Wheeler's thesis on *An Analysis of Tests of Manual Dexterity* (Univ. London Library) which I hope may shortly be published.

[4] In early memoranda I suggested graphing correlations *after the table had been rearranged* (*a*) to determine how far the table was 'synclinal' ('hierarchical') or 'heteroclinal,' and (*b*) to estimate the self-correlations. The idea was borrowed from Karl Pearson, who called such graphs 'analographs.' The procedure presents an interesting picture of the relations suggested; but it was found to be too lengthy and inaccurate to form a regular part of the routine procedure.

way of partitioning the table, because the profiles and the peaks of the observed correlations may be affected both by the general factor and by the errors of sampling.

(4) *Intercolumnar Correlations.* Spearman, it will be remembered, in contrasting the older hypothesis of 'multifocal levels' with his own 'unifocal' hypothesis, suggested, as a crucial test, the value of the correlation between the various columns of observed correlations. "When the columns belong to [tests from] different levels, the high coefficients would always come in non-corresponding places, so that the correlation between columns is *negative*" (*Brit. J. Psych.*, V, p. 57; cf. 6, p. 139). With my own table, which he reprints, he finds the (uncorrected) intercolumnar correlation to be, for the four largest cases, + 0·81, for the five smallest + 0·59. He therefore infers that, contrary to my own contention, the table gives no evidence whatever for group factors. Miss Carey, however, has pointed out that, unless the amount of correlation due to the group factors is large and the amount due to the general factor comparatively small, the intercolumnar coefficients will *not* be negative, but only reduced in size. In practice therefore, particularly if there is some measure of overlapping, the criterion may often prove somewhat inconclusive.

(5) *Preliminary Bipolar Analysis.* The method which to my mind is by far the best for general purposes is to start with an ordinary factor analysis by simple summation, and then to take the positive and negative sections of the bipolar factors (provided their saturations are significant) as indicating where the lines of partition are to be drawn. This procedure is entirely automatic and objective. In certain cases a preliminary analysis of this kind may be dispensed with: when the correlations have been rearranged so as to bring clusters of high coefficients as near to the diagonal as possible, it will sometimes be found that the lines of classification are sufficiently clear without further check. But such cases are exceptional: whenever there is the smallest room for doubt, an initial bipolar analysis is in my view indispensable.

The Addition and Multiplication Procedures. In the group factor method, as compared with other forms, the chief novelty arises from the way the first or 'general' factor has to be computed. Since it is not precisely the same as the 'general' factor reached by ordinary 'simple summation,' it is perhaps better to give it a distinctive name. As I have stated elsewhere, the underlying aim is to "seek a general factor which will represent, not the *average* slope or plane of the correlations, but the *basic* slope" (7, p. 308). Accordingly, we may call the factor so obtained the 'basic' factor. Our discussion will therefore be concerned chiefly with the different devices available for calculating this factor.

With the group factor hypothesis, as we have seen, the essential assumption is that all the factors except the first are limited (save for a sporadic overlap) to groups of 'like' tests; and that in consequence *those parts of the correlation table which relate to 'unlike' tests will form parts of a 'hierarchy'* (matrix of rank one), since they are due solely to the single basic factor (except for occasional overlaps which we have agreed to consider later). Now, in an ideal case, where the correlations in this portion of the table are *exactly* hierarchical, it is easy to show[1] that the fundamental equation given above yields two alternative procedures for calculating the saturations, which in such a case will furnish precisely the same figures. The first requires us to *add* the row-totals; the second to *multiply* them.

The multiplication method leads to a formula which may at first sight seem simpler; and it forms the basis of the revised form of Holzinger's 'bifactor method.'[2] On the other hand, if I may trust my own experience, what I have called the addition method appears preferable in actual practice. In my original memorandum two reasons were briefly mentioned for this preference. First, where the observed coefficients (owing it may be to large fluctuations due to sampling) depart appreciably from a strict hierarchical arrangement, there the multiplication method is apt to exaggerate the effect of these departures : in certain cases it may even give saturations over unity ; in others, where one of the sums to be multiplied is nearly zero, it may give saturations that are too near to zero. Secondly, the multiplication method does not lend itself to the usual type of check. In view of the questions that have recently been raised, these points perhaps deserve exemplification at greater length.

---

[1] To save space I shall not give algebraic proofs for all the formulæ implied by the procedures here described. The more important are outlined in the Appendix below, and are discussed in greater detail in the roneo'd *Laboratory Notes on Factor Analysis: Group Factor Method.* There the fuller exposition gives, for each procedure, (i) the algebraic derivation of the method; (ii) a simple illustration, based on an artificial correlation table, exemplifying the essential principle; (iii) an application to an empirical correlation table. From the discussion and references below, the interested reader will, I imagine, have little difficulty in supplying more details for himself, and finding appropriate examples in the literature.

[2] Holzinger, K. J. (1937). *Student Manual of Factor Analysis*, pp. 13f. As originally described, the 'bifactor method' assumed that the common factors could be "thought of as split into two *proportional* parts, one of which is the principal factor." As a consequence, the saturation coefficients given for the secondary factors in the illustrative tables were always *exactly* proportional to the corresponding coefficients for the principal factor: cf. Holzinger, K. J. (1935), *Preliminary Reports on the Spearman-Holzinger Unitary Trait Study*, No. 5, p. 6 and table on p. 2. This conception was originally put forward by Spearman, who sometimes called it a 'bifid factor.' It is referred to by several of Spearman's students, e.g., Hargreaves and Stephenson, but was not, I think, actually adopted in any of their investigations. The earliest published reference to it is in the appendix to H. L. Hargreaves' monograph, where he discusses the alternative forms which Spearman's theory

A. THREE GROUP FACTORS

Example. May I begin with what is at once the simplest and the commonest type of case, namely, that in which there are three group factors? Most of the worked examples given to illustrate this or allied procedures are based on tables of artificial correlations which give exact results. In such cases the figures obtained with Holzinger's method and my own are likely to be identical. To illustrate the differences it is therefore desirable to choose a table that deviates to some extent from the exact hierarchical figures. To keep the working as simple as possible, let us take a table involving two tests only in each group (Table I). As usual, the chief difficulty arises, not over the calculation of the group factors, but over the extraction of the basic factor. I shall therefore omit the square submatrices affected by the group factors.

TABLE I. CALCULATION OF SATURATIONS FOR THE BASIC FACTOR

Matrix of Observed Correlations, Partitioned and Condensed

Tests	1	2	Sum	3	4	Sum	5	6	Sum
1	—	—	—	·40	·26	·66	·44	·30	·74
2	—	—	—	·55	·38	·93	·05	·02	·07
Sum ...	—	—	—	·95	·64	1·59(*A*)	·49	·32	·81(*B*)
3	·40	·55	·95	—	—	—	·18	·10	·28
4	·26	·38	·64	—	—	—	·11	·09	·20
Sum ...	·66	·93	1·59(*A*)	—	—	—	·29	·19	·48(*C*)
5	·44	·05	·49	·18	·11	·29	—	—	—
6	·30	·02	·32	·10	·09	·19	—	—	—
Sum ...	·74	·07	·81(*B*)	·28	·20	·48(*C*)	—	—	—
Grand Total	1·40	1·00	2·40	1·23	·84	2·07	·78	·51	1·29
Divisors ...	1·4652			2·1325			2·6087		
Sat. (Add.)	·9555	·6825	1·6380	·5768	·3939	·9707	·2990	·1955	·4945
Sat. (Mult.)	1·0087	·3682	—	·5731	·3975	—	·2989	·1956	—

With both the addition and the multiplication procedure the work is considerably facilitated if the sub-totals for the row-sections and column-sections are calculated at the outset, and entered in the table of observed correlations as shown above ; the totals for the submatrices and for the

---

of a general factor might take : among the conditions under which " a general factor can be compound, and so act as a single general factor," one is that in which " two (or more) factors enter into every test in exactly the *same proportion* " (*Brit. J. Mon. Sup.*, X, 1927, p. 66 : author's italics). Stephenson's 'unanalysed abilities,' due to 'non-fractional factors,' are of the same type (*Brit. J. Psych.*, XXX, 1937, pp. 100–102).

entire columns can be calculated from these, and entered at the same time. This method of setting out such tables was suggested in the working instructions for my own procedure (e.g., Table VIII, *Factors of the Mind*, p. 478); but it can also be adopted to illustrate the method proposed by Holzinger (cf., for example, Thomson, 11, table on p. 339).

1. *The Addition Method*. As in applying the principle of simple summation to bipolar analysis, so here the addition method requires us, after adding each column, to divide their totals by an appropriate divisor. The quotients will then give us the saturation coefficients for the first factor. With group factor analysis, however, we have to omit the diagonal submatrices, since these depend on group factors as well as on the general factor. This omission will naturally entail a corresponding modification in the mode of calculating the divisors. Instead of one divisor for all the column totals (as in bipolar analysis) we shall have to use as many different divisors as there are omitted submatrices—in the present example, three.

Let us use the letters A, B, and C to denote the subtotals of the three non-diagonal submatrices, as indicated in the table. Then, as may be easily shown (see Appendix, p. 73), for the first two columns (i.e., for the coefficients entering into the submatrices A and B) the formula for the divisor will be $\sqrt{C}\left\{\sqrt{\frac{A}{B}}+\sqrt{\frac{B}{A}}\right\}=\sqrt{C}\left\{\frac{A+B}{\sqrt{AB}}\right\}$. It will be observed that the figures required for the terms inside the braces are the sums of those submatrices whose column totals are to be divided. Thus, in the case of the two columns comprising submatrices A and B, the divisor will be $\sqrt{\cdot 48}\left\{\sqrt{\frac{1\cdot 59}{\cdot 81}}+\sqrt{\frac{\cdot 81}{1\cdot 59}}\right\}=1\cdot 4652$. Similarly, for the next two columns, it will be $\sqrt{\cdot 81}\left\{\sqrt{\frac{1\cdot 59}{\cdot 48}}+\sqrt{\frac{\cdot 48}{1\cdot 59}}\right\}=2\cdot 1325$; and for the last two, $\sqrt{1\cdot 59}\left\{\sqrt{\frac{\cdot 81}{\cdot 48}}+\sqrt{\frac{\cdot 48}{\cdot 81}}\right\}=2\cdot 6087$. On dividing each column total by the appropriate divisor as just computed, we obtain the first row of saturations shown in the last line but one.

The theoretical matrix of rank one, to be fitted to the matrix of observed correlations, is obtained in the usual way by multiplying each saturation coefficient by all the rest. This is analogous to the device first proposed by Pearson for fitting a theoretical matrix of rank one to an observed 'contingency table' in cases of 'manifold association' (cf. Yule (1910), *Introduction to Statistics*, p. 64, eq. 1, and refs.). To test goodness of fit we may accordingly adopt the chi-squared procedure, employed by Pearson for such cases. The residuals are found by subtracting each theoretical correlation from the corresponding observed correlation; then squared, summed, and multiplied by the number of persons tested. This was the method adopted in earlier investigations both with bipolar and with group factor analysis. However, in the latter case we need to locate the pairs of tests which appear to indicate an overlap of factors. Hence for most purposes it is simpler and more effective to test each residual separately by the standard error of the observed correlation. If there is no overlap, the residuals in the three side submatrices should *all* be non-significant.[1]

Checks.—(i) With this procedure, as a little algebra will show,[2] each column of residuals

---

[1] The only significance test that could be rigorously defended on statistical grounds would require the theoretical values to be fitted by the method of least squares. We could then derive a likelihood ratio from the products of the factor variances or latent roots. But the requisite formula would, I fancy, be too complex for practical use. Mr. J. R. Robertson has, at my suggestion, applied the simpler significance-tests to artificial tables constructed by using Tippett's random numbers, and finds those described in the text to be reasonably satisfactory, provided the observed correlations are not large and the number of persons tested (N) is not small. When the observed correlations are large, I have recommended one or other of two methods of correction. In earlier investigations the residuals were augmented by using Yule's formula for partial correlation. In later investigations, carried out after 1925, we have commonly used Fisher's z-transformation, and multiplied the square-sum by $(N - 3)$ instead of N. (This method of testing discrepancies between correlations has also been used for more general purposes, e.g., by Tippett himself, *Methods of Statistics*, p. 144.)

[2] Proofs of all these checks are given in *Laboratory Notes* cited above.

(excluding, of course, the residuals in the diagonal submatrices) should add up to zero. We can thus apply the same kind of check that we use in other forms of summational analysis. By way of illustration I have calculated, in Table II below, the residuals for the first column.

TABLE II. ADDITION METHOD. CHECKING THE RESIDUALS

Tests	Correlations		Residuals
	Observed	Calculated	
1 and 3 … …	·40	·5511	− ·1511
4 … …	·26	·3764	− ·1164
5 … …	·44	·2857	+ ·1543
6 … …	·30	·1868	+ ·1132
Sum … …	1·40	1·4000	·0000

(ii) The second check is also analogous to that used in simple summation. (*a*) If we add all the saturations, except those for a particular block, their sum should be equal to the divisor for that block. Thus the sum of the first four saturations is 1·6380 + ·9707 = 2·6087, which is identical with the last divisor (see Table I). (*b*) This check may be put in a slightly different form. If, instead of adding the subtotals for the saturations, we multiply them, their product should be equal to the total of the submatrix common to the two blocks involved. Thus 1·6380 × ·9707 = 1·59 = *A*. Similarly with the other sets of saturations.

2. *The Multiplication Method.* With only three factors the multiplication method reduces to a procedure [1] which, at first sight, seems even simpler than the addition method. Instead of adding the row- (or column-) sums of the separate submatrices, we multiply them: e.g., for test 1 (which enters into submatrices *A* and *B*), instead of taking the sum (·66 + ·74) = 1·40, and then dividing it by a divisor calculated as above, we take the product of those figures (·66 × ·74) = ·4884, and divide this product by the sum of the remaining submatrix (*C*), namely, ·48. The saturation coefficient is the square root of this quotient. The calculation of the several saturations thus proceeds as follows. Test 1, $\sqrt{(\cdot 66 \times \cdot 74) \div \cdot 48} = 1\cdot 0087$; test 2, $\sqrt{(\cdot 93 \times \cdot 07) \div \cdot 48} = \cdot 3682$; test 3, $\sqrt{(\cdot 95 \times \cdot 28) \div \cdot 81} = \cdot 5731$; etc.

We note (i) that the saturation for the first test here rises to over 1·0000, an impossible figure, and (ii) that the saturation for the second test falls to an unduly low figure, ·3682. Generally speaking, it will be seen that, wherever owing to errors of sampling the observed coefficients happen to be too large (or too small) for a precise hierarchical fit, the process of multiplication tends to enlarge (or to diminish) the resulting saturation to an exaggerated degree.

Neither of the checks described above is applicable here. Thus the residuals for the first test are as follows:

TABLE III. MULTIPLICATION METHOD. CHECKING THE RESIDUALS

Tests	Correlations		Residuals
	Observed	Calculated	
1 and 2 … …	·40	·578	− ·178
3 … …	·26	·401	− ·141
4 … …	·44	·302	+ ·138
5 … …	·30	·197	+ ·103
Sum … …	1·40	1·478	− ·078

---

[1] Holzinger, I think, does not explicitly illustrate how his own method is to be used in the case of three group factors only. The steps required, however, can be readily deduced from his procedure (described below) for the case of more than three factors; but with three factors only, the actual working becomes greatly simplified. See Holzinger's *Student Manual*, p. 26.

It will also be noted that the saturations given by my own procedure yield a much closer fit. The mean discrepancy (i.e., the average residual) is ·092 with the addition method, and ·107 with the multiplication method.

B. MORE THAN THREE GROUP FACTORS

1. *The Addition Method.* The principle described in section A1 above can readily be extended to the case in which more than three group factors are required.[1] But, as in factorizing individual correlations, so in factorizing the sums obtained by condensation : whenever there are more than three given figures to be factorized, i.e., when we are concerned with a square matrix of an order larger than 3 × 3, we must insert estimated values for the missing cells in the leading diagonal.

As an illustrative example let us take the following table. To avoid introducing large decimal fractions, which would not only take up undue space but also be harder to follow, I have chosen artificial figures ; and the values for the ' observed ' correlations are made to differ very slightly from those which would have been obtained by multiplying the saturation coefficients assumed when constructing the table.

As before, we first rearrange, partition, and condense the matrix of individual correlations (Table IV). The row- and column-sums, and the totals for each column, are then calculated, as for Table I.

TABLE IV. CALCULATION OF SATURATIONS FOR THE BASIC FACTOR

Matrix of Observed Correlations Partitioned and Condensed

Test	1	2	(*a*) Sum	3	4	(*b*) Sum	5	6	(*c*) Sum	7	8	(*d*) Sum
1	—	—	—	·63	·55	1·18	·44	·36	·80	·27	·18	·45
2	—	—	—	·57	·48	1·05	·40	·32	·72	·23	·16	·39
(*a*) Sum ...	—	—	—	1·20	1·03	2·23	·84	·68	1·52	·50	·34	·84
3	·63	·57	1·20	—	—	—	·35	·27	·62	·21	·14	·35
4	·55	·48	1·03	—	—	—	·30	·24	·54	·18	·11	·29
(*b*) Sum ...	1·18	1·05	2·23	—	—	—	·65	·51	1·16	·39	·25	·64
5	·44	·40	·84	·35	·30	·65	—	—	—	·16	·10	·26
6	·36	·32	·68	·27	·24	·51	—	—	—	·12	·09	·21
(*c*) Sum ...	·80	·72	1·52	·62	·54	1·16	—	—	—	·28	·19	·47
7	·27	·23	·50	·21	·18	·39	·16	·12	·28	—	—	—
8	·18	·16	·34	·14	·11	·25	·10	·09	·19	—	—	—
(*d*) Sum ...	·45	·39	·84	·35	·29	·64	·26	·21	·47	—	—	—
Grand Total	2·43	2·16	4·59	2·17	1·86	4·03	1·75	1·40	3·15	1·17	·78	1·95
Divisors ...	2·70			3·10			3·50			3·90		
Saturations	·90	·80	1·70	·70	·60	1·30	·50	·40	·90	·30	·20	·50

We have now to factorize the sums of the submatrices. This involves factorizing a condensed 4 × 4 matrix instead of the original 8 × 8 matrix. The sums, already calculated in the body of Table IV, are for convenience set out afresh in Table V. To factorize Table V by simple summation, we must first complete the table by inserting suitable figures in the diagonal. These, of course, should be the squares of the saturations we are about to calculate. But, as we do not yet know the saturations, we must proceed by successive approximation; in this way we shall ultimately reach the figures shown in the diagonal within brackets. The calculation of the saturations is then accomplished in the usual manner, by dividing each column total by the square root of the grand total. (It will be noted that the figures inserted in brackets agree with the squares of the saturations so found: $2{\cdot}89 = 1{\cdot}70^2$, etc.)

---

[1] In describing my procedure Thomson observes that " Burt only gives his formula for three groups " (11, p. 341). However, in my paper on ' Factor Analysis by Submatrices,' the use of the method with more than three groups was briefly described and illustrated (8, pp. 350f., section headed " *More than three group factors*," and equation iv, p. 352 ; cf. also 9, p. 482, note in small print). It had previously been employed in one or two L.C.C. investigations and in several unpublished theses by research students.

TABLE V. ADDITION METHOD

Calculation of Divisors from Condensed Correlation Matrix

Group	Sum of Submatrices (*a*)	(*b*)	(*c*)	(*d*)	Total
(*a*)	(2·89)	2·23	1·52	·84	7·48
(*b*)	2·23	(1·69)	1·16	·64	5·72
(*c*)	1·52	1·16	(·81)	·47	3·96
(*d*)	·84	·64	·47	(·25)	2·20
Divisor 4·4	7·48	5·72	3·96	2·20	19·36 = 4·4
Sats. for Sums ...	1·70	1·30	·90	·50	4·40
Reference letter ...	S_a	S_b	S_c	S_d	T
Divisors	2·70	3·10	3·50	3·90	13·20

Let us call the 'saturations for sums' S_a, S_b, S_c, and S_d respectively, and their total T. Then the divisor for the columns on which S_a is based will be $T - S_a$; for those on which S_b is based, $T - S_b$; and so on. The total of the divisors $= (m - 1)T = 3 \times 4{\cdot}40 = 13{\cdot}20$, where m as before denotes the number of groups (here four). On dividing the totals for the test-columns, as given in the initial table (Table IV), we obtain the eight saturation coefficients shown in the bottom line: thus $2{\cdot}43 \div 2{\cdot}70 = {\cdot}90$, etc.

Checks. The same two checks can be used as before: viz., (1) the sum of all the saturations except those belonging to a given block should be identical with the divisor for that block; (2) the residuals should add up to zero.

2. *The Multiplication Method*. As we have seen, the bifactor method depends essentially on the multiplication of sums rather than on the addition of sums. The working procedure may be summarized as follows.[1]

First, the totals of the sums of the submatrices are found as in Table V, but *without* inserting estimated figures for the diagonal cells: thus, $2{\cdot}23 + 1{\cdot}52 + {\cdot}84 = 4{\cdot}59$. These totals are as a matter of fact given at the foot of the columns headed (*a*), (*b*), (*c*), and (*d*) in Table IV. But to make the working quite clear, the sums have been copied out and totalled afresh in Table VI. To find the divisors, the totals so obtained are subtracted from *half* the grand total ($6{\cdot}86 - 4{\cdot}59 = 2{\cdot}27$, etc.).

TABLE VI. MULTIPLICATION METHOD

Calculation of Divisors for Bifactor Analysis

Group	Sum of Submatrices (*a*)	(*b*)	(*c*)	(*d*)	Total
(*a*)	—	2·23	1·52	·84	4·59
(*b*)	2·23	—	1·16	·64	4·03
(*c*)	1·52	1·16	—	·47	3·15
(*d*)	·84	·64	·47	—	1·95
Total	4·59	4·03	3·15	1·95	13·72 ÷ 2 = 6·86
Divisors	2·27	2·83	3·71	4·91	13·72

---

[1] See 10, Appendix C, pp. 328–341.

A second supplementary table (Table VII) is now constructed from Table IV as follows. The first figure in each row of Table VII consists of the total of all the sums in the corresponding row of Table IV, except the first sum (·80 + ·45 = 1·25); the next figure of the total of all the sums except the first and second; and so on if there are more totals left to calculate. In the present case the second 'total' is simply the last figure in the corresponding row of Table IV, i.e., ·45. It is convenient to enter each of these 'diminishing sums' under the letter corresponding to the omitted sum.

TABLE VII.

TABLE OF DIMINISHING SUMS

Group	(*a*)	(*b*)	(*c*)
Test 1	—	1·25	·45
2	—	1·11	·39
3	·97	—	·35
4	·83	—	·29
5	·91	·26	—
6	·72	·21	—
7	·67	·28	—
8	·44	·19	—

TABLE VIII.

PRODUCTS FOR NUMERATOR

Group	(*a*)	(*b*)	(*c*)	Sum
Test 1	—	1·4750	·3600	1·8350
2	—	1·1655	·2808	1·4463
3	·1640	—	·2170	1·3810
4	·8549	—	·1566	1·0115
5	·7644	·1690	—	·9334
6	·4896	·1071	—	·5967
7	·3350	·1092	—	·4442
8	·1496	·0475	—	·1971

These 'diminishing sums' have now to be multiplied by the figures that were successively omitted, i.e., the first and second subtotals from Table IV (the last subtotal, for *d*, will not be required, because there is no entry corresponding to it in Table VII [1]). We obtain the products shown in Table VIII. The products in each row are then summed; and the sums divided by the group divisors already found in Table VI. The square root of the quotient gives the saturation for the test (see Table IX). It will be remarked that, although the 'observed' correlations differ only very slightly from the exact hypothetical values, as constructed from the saturation assumed, the bifactor method does not reproduce those saturations so accurately as the group factor method ('addition method'). Further, the checks are no longer applicable.

TABLE IX. CALCULATION OF SATURATIONS

Test	Sum of Products (Numerator)	Divisor (Denominator)	Quotients	Saturations (Square Root)	True Values
1	1·8350	2·27	·8084	·8991	·90
2	1·4463	2·27	·6371	·7982	·80
3	1·3810	2·83	·4880	·6986	·70
4	1·0115	2·83	·3574	·5978	·60
5	·9334	3·71	·2516	·5015	·50
6	·5967	3·71	·1609	·4011	·40
7	·4442	4·91	·0905	·3008	·30
8	·1971	4·91	·0401	·2003	·20

Comparison of Alternative Methods. The theoretical grounds for preferring what I have called the addition method will, I hope, now be clear. With the addition method we are in effect basing our estimate of the saturation coefficient on a process of averaging. In fact, the expression for the saturation coefficient, with this mode of calculation, might be regarded as a kind of a weighted average, the divisor being, as usual, the sum of the weights. This is the natural principle to adopt in smoothing

---

[1] The whole process is equivalent to multiplying each sum by the total of all the rest, and then halving the sum of the products. The 'total of all the rest' can be obtained by deducting in turn each sum from the column total. Thus for the first row we should have the following figures:

		(*a*)		(*b*)		(*c*)		(*d*)		Sum
Sums	...	—	...	1·25	...	1·63	...	1·48		
Products	...	—	...	1·4750	...	1·3040	...	0·8910	...	3·6700 = 2 × 1·8350

out the effects of the sampling fluctuations that must affect each observed coefficient. The multiplication formula, on the other hand, is a direct algebraic deduction from a postulated equation ; and the deduction is strictly applicable only when that postulated equation is exactly fulfilled. With an empirical table this can never be the case. Consequently, the method is exposed to all the risks incurred when errors are implicitly multiplied instead of being added.[1]

A few words are perhaps necessary to indicate how the procedures so far described differ from that adopted by Spearman and his followers. There are three points of divergence which deserve some comment. (i) The principal difference arises from the fact that Spearman always distrusted any attempt to assess the general factor on the basis of internal evidence. He therefore preferred to use ' objective tests ' of ' pure *g*,' and to eliminate *g* by the ordinary formula for partial correlation. This would seem to be a perfectly legitimate procedure, provided (*a*) that the general factor we wish to eliminate is identifiable with *g* as independently assessed (i.e., with what is ordinarily described as ' general intelligence '), and (*b*) that we possess a satisfactory set of tests for assessing ' pure *g*.' Otherwise it would seem wiser to take the information supplied from the entire battery of tests into account. (ii) But further, Spearman himself thought it necessary, on methodological grounds, to deal with only one group factor at a time. Thus, in re-examining the evidence for the various group factors adduced by other investigators, instead of analysing an entire correlation table which might need to be partitioned for three or more group factors, he preferred to confine himself to a table, or part of a table, which could be reduced to a 2 × 2 form, consisting of inter-correlations and cross-correlations between two groups or tests only. And here, instead of looking for a group factor in each of the two groups, he insisted, as before, on treating one of the groups as consisting of ' reference values,' dependent solely on *g*, while the second group was examined for the presence of a possible group factor. (iii) Thirdly, Spearman objected to my attempt to derive group factors by summing an entire submatrix of ' like ' tests for the same reasons that he objected to my attempt to derive a general factor by summing the entire correlation table. In both cases factors were treated as averages ; and in his view this was a mistake (6, pp. 61f.). Strictly speaking, his own formula was applicable only to the case in which there were *two* ' reference abilities ' and in which *two* ' abilities ' were to be examined for specific correlation (6, p. 223, and Appendix, p. xxi). Consequently, " some complication," as he says, " is introduced when the available abilities are more than two in number, for not more than two can be used at the same time. . . . The more thorough way is to try out every possible pair." [2]

His method of looking for a single group factor only I shall discuss more fully in a moment. Meanwhile, let us consider how the group factor method can be adapted to the case where there are only two group factors.

C. TWO GROUP FACTORS

Frequency of Paired Group Factors. The need to deal with two group factors occurs in three main types of investigation.

(i) It occurs in dealing with those correlation tables that are reducible to a 2 × 2 form, with which, as we have just seen, Spearman preferred to work. The procedure that he proposed to substitute is by no means satisfactory : few of his critics accepted his demand that, in every given analysis, all group factors but one should be dispensed with. Indeed, more often than not, the tests which he employed to eliminate *g* seemed themselves to involve a second group factor of their own. For example, in dealing with Miss Rogers' study of mathematical ability, he reduced the tests

[1] There are other differences between the two methods : e.g., Holzinger's regression equation for estimating factor-measurements (*Preliminary Reports*, No. 8, p. 7) includes only those tests that contain the group factor ; mine uses all the tests for all the factors. I have discussed this and other differences more fully in (8). As stated above, full proofs of the relevant formulæ are given in the roneo'd *Notes* already cited. But once again, I imagine, the reader will have no difficulty in working them out for himself. In my notes, instead of the notation of the Appendix, I suggested writing $a_1, a_2, \ldots b_1, b_2, \ldots$ etc., for the hypothetical saturations in groups *a*, *b*, etc. The expressions for the working equations and their proofs can then be readily written down.

[2] In practice, however, he considered it both legitimate and sufficient to " pool together all those abilities that enter into the same set." But in principle there seems very little difference between the proposal to " pool " and the proposal to calculate factors by " summation " or " averaging." I should add, however, that the formula which Spearman suggested for use when pooling was not altogether correct, even if his own assumptions are accepted (see this *Journal*, I, pp. 96f.).

employed to two main groups : (*a*) ' arithmetical tests,' to be examined for a possible group factor, and (*b*) ' reference values,' to serve as tests of pure *g*. Now the latter included Analogies, Opposites, Completion, and Thorndike's Test of Reading. All of them were verbal tests ; and it seemed highly likely that they contained a verbal factor as well as ' pure *g*.' The same was true of his discussion of Collar's research—a somewhat similar study of the arithmetical factor based on work in L.C.C. schools : here again it seemed clear that each of the two contrasted groups of tests included its own specific group factor.[1]

(ii) In the analysis of physical traits a similar situation often occurred. In the earliest empirical table which seemed to call for an analysis in terms of group factors—namely, the Pearson-Macdonell table for head and body measurements to which I have already referred—the requisite division was clearly into two main clusters of ' like traits.' The procedure which I very tentatively proposed drew sharp criticisms both from Spearman and from Pearson himself. In later work with group factors, therefore, to avoid objections directed primarily against the technique adopted rather than against the conclusions drawn, we generally tried to arrange the battery of tests or traits so that there should, if possible, always be at least three group factors. In psychological work, it was quite easy to secure this condition : indeed, in early work with mental tests, where the number of persons tested was small, and the analysis could not pretend to isolate more than a couple of factors in each test, this threefold grouping was the type most commonly encountered. On the other hand, in the field of physical measurement, the study of ' body types '—particularly of the recurring contrast between the leptomorphic (or asthenic) type and the eurymorphic (or pyknic) type—constantly called for analysis in terms of two group factors.

(iii) Later, when larger populations became available for mental testing, it was often found that many of the group factors hitherto discovered were apt to get subdivided in their turn ; and the division was usually by pairs. Indeed, an extremely common type of factor-pattern proved to be that in which the whole set of traits involves two broad group factors, and then each of these splits into two subfactors, and so on. Hence it seemed urgently necessary to devise an acceptable method which should at once be convenient in practice and free from serious theoretical objections.

Now, even those investigators who have attempted to deal with several group factors in the same correlation table appear generally to assume that, unless a battery includes at least one test which involves no group factors at all but only the general factor, it is impossible to determine the factor-saturations with two groups only : just as " three variables at least are always necessary for a Spearman ' two factor ' analysis," so (it seems to be supposed) three groups at least are always necessary for a group factor analysis. Thus, in illustrating how a set of seven tests can be factorized to give one general and two group factors, Holzinger constructs his artificial correlation table so that the seventh test has zero saturations for every factor except the first.[2] Actually, however, there are several ways in which the difficulty can be overcome. Here I shall content myself with what in practice has proved to be the simplest and most convenient.

Example. In contributions previously published I have always applied the method to empirical figures obtained from actual researches. But for purposes of illustration, the errors and discrepancies inevitably involved in dealing with observed coefficients tend to obscure the underlying principles. Here, therefore, I shall take an artificial table of correlations, constructed to give simple and exact calculations. A convenient set of figures, obtained by rounding off observed results from tests of educational abilities,[1] is shown in Table X. The coefficients, as we shall see in a moment, have been built up by using one basic and two non-overlapping group factors; the table therefore has a rank of three. Accordingly, by successive approximation, a computer, given nothing but the test-correlations, should be able to determine exactly the appropriate values for the self-correlations (shown in Table X within brackets).

---

[1] Rogers, A. (1923), ' Tests of Mathematical Ability,' Columbia Univ. Cont., No. 130, Table IX, pp. 56–57 ; Collar, D. J. (1920), ' A Statistical Survey of Arithmetical Ability,' *Brit. J. Psych.*, XI, pp. 135f. Miss Rogers' data seemed to me to afford admirable support for my contention. Actually it included a *third* group of tests, mainly perceptual, which Spearman omitted. It was thus possible to show that the verbal factor, which appeared with my ' two-group-factor analysis ' of the small table of correlations selected by Spearman, was identical with the verbal factor which appeared with a ' three-group-factor analysis,' when the *whole* of Miss Rogers' table was employed. With two group factors the chief objection urged by Spearman and others was that the factors and their saturations could not be uniquely determined. But with three group factors, of course, there could be no doubt about the uniqueness of the figures obtained. Spearman, however, further maintained that, in any case, no such thing as a group factor for verbal ability existed (cf. *Abilities of Man*, p. 237). In later editions of his book he altered his view, and added this factor to his list. The reader who desires to try out the methods described in the remainder of this article will find in Miss Rogers' figures an empirical table admirably adapted for the purpose.

[2] Holzinger, H. *Student Manual*, p. 26.

TABLE X. OBSERVED CORRELATIONS

Test	1	2	3	4	5	6	7
1. Verb. Int.	(·82)	·76	·48	·29	·63	·45	·36
2. Sci. 1 (Gen.)	·76	(·80)	·52	·32	·56	·40	·32
3. Sci. 2 (Mech.)	·48	·52	(·34)	·21	·35	·25	·20
4. Dict.	·29	·32	·21	(·13)	·21	·15	·12
5. Non-Verb. Int.	·63	·56	·35	·21	(·53)	·43	·32
6. Mech. Ass.	·45	·40	·25	·15	·43	(·41)	·28
7. Sci. 3 (Pract.)	·36	·32	·20	·12	·32	·28	(·20)

TABLE XI. FACTOR-SATURATIONS

Tests	A. Bipolar Analysis				B. Group Factor Analysis			
	I	II	III	Square-Sum	*G*	*V*	*P*	Square-Sum
1. Verb. Int. ...	·8824	+·0625	+·1939	·8200	·90	·10	—	·82
2. Sci. 1 (Gen.) ...	·8567	+·2500	−·0596	·8000	·80	·40	—	·80
3. Sci. 2 (Mech.) ...	·5471	+·1874	−·0747	·3400	·50	·30	—	·34
4. Dict.	·3329	+·1249	−·0596	·1300	·30	·20	—	·13
Sum (Verbal) ...	2·6191	+·6248	·0000	—	2·50	1·00	—	—
5. Non-Verb. Int....	·7054	−·1562	+·0894	·5300	·70	—	·20	·53
6. Mech. Ass. ...	·5518	−·3124	−·0894	·4100	·50	—	·40	·41
7. Sci. 3 (Pract.) ...	·4190	—·1562	·0000	·2000	·40	—	·20	·20
Sum (Practical) ...	1·6762	−·6248	·0000	—	1·60	—	·80	—
Square-Sum ...	2·9003	·2635	·662	3·2300	2·69	·30	·24	3·23
Contrib. to Variance (per cent.)	41·4	3·8	·9	46·1	38·4	4·3	3·4	46·1

As usual we begin with a preliminary 'bipolar analysis.' The saturations for the general and two bipolar factors, found in this way, are shown in Table XI A. The figures indicate that a clear line of division can be drawn between the first four tests and the last three. As with the Pearson-Macdonell table, the cross-correlations between the four traits forming the first group and the three traits forming the second group fit a hierarchical arrangement. These cross-correlations are shown as a separate table in Table XII. Since they form a matrix of rank one, they can be regarded as due to a single common factor only, namely, the 'basic' factor.

---

[1] The tests are described below on p. 62.

(i) *Determining the Saturations for the Basic Factor.* We now proceed to analyse this rectangular submatrix in much the same way as we should analyse a complete square matrix of observed correlations, when extracting a single factor only. Since, however, the submatrix is asymmetrical, the sums of the columns will no longer be identical with the sums of the rows. We must therefore consider how we can find a suitable divisor. To take the square root of the grand total, as we do with a symmetrical table, would, as a general rule, be quite indefensible. For one thing, there are four tests furnishing column-sums, but only three furnishing row-sums; and this fact of itself must make the column-sums smaller. Secondly, quite apart from the difference in the number of tests in either group, we have no right to assume that the factor-saturations in each group will be equally balanced: as we shall find in a moment, the average saturation for the first three tests is only 0·53, and for the last four 0·63.

To allow for these two peculiarities we must therefore use two divisors instead of one. For a first approximation the simplest procedure is to select the divisors so that in the group factor analysis the sum of the factor-saturations for the first three tests will have the same proportion to the sum of the factor-saturations for the last four tests as they do in the bipolar analysis.

Now let d_1 and d_2 be the two divisors required for calculating the group factor-saturations from the row- and column-sums; t_1 and t_2 the sums of the group factor-saturations for the corresponding groups; t the sum of all the correlations in the submatrix; and s_1 and s_2 the sums of the factor-saturations obtained for the first and second groups of tests by the preliminary bipolar analysis. Then

$$t_1 = \frac{t}{d_1}, \text{ and } t_2 = \frac{t}{d_2}.$$

But $t = t_1 t_2 = d_1 d_2$. Hence $d_1 = t_2$, and $d_2 = t_1$.

Now, since $\frac{d_2}{d_1} = \frac{t_1}{t_2} = \frac{s_1}{s_2}$, we have $d_2 = \frac{s_1}{s_2} d_1$; and $t = d_1 d_2 = \frac{s_1}{s_2} d_1^2$.

Hence $$d_1 = \sqrt{\left\{\frac{s_2}{s_1} t\right\}} \text{ and } d_2 = \sqrt{\left\{\frac{s_1}{s_2} t\right\}}.$$

Thus all we have to do is to multiply the total (t) by the proportion of the two sums as given by the bipolar analysis (s_1/s_2 or s_2/s_1), and take the square root. Accordingly, inserting the appropriate figures from Tables XI A and XII, we find

$$d_1 = \sqrt{\left\{\frac{2{\cdot}6191}{1{\cdot}6762} \times 4{\cdot}00\right\}} = 2{\cdot}50,$$

$$\text{and } d_2 = \sqrt{\left\{\frac{1{\cdot}6762}{2{\cdot}6191} \times 4{\cdot}00\right\}} = 1{\cdot}60.$$

On dividing the row-sums by 2·50 and the column-sums by 1·60, we obtain the saturations shown on the extreme right and at the foot of Table XII. We can check the figures along the same lines as before by noting that the sum of the saturations for tests 1 to 3, viz., $t_1 = 1{\cdot}60 = d_2$; and the sum for tests 4 to 7, namely, $t_2 = 2{\cdot}50 = d_1$. Further, $1{\cdot}60 \times 2{\cdot}50 = 4{\cdot}00$.

TABLE XII. CALCULATION OF SATURATIONS FOR BASIC FACTOR (G)

Non-Verbal Tests	Verbal Tests 1	2	3	4	Row-Sums	Sats.
5	·63	·56	·35	·21	1·75	·70
6	·45	·40	·25	·15	1·25	·50
7	·36	·32	·20	·12	1·00	·40
Column-Sums	1·44	1·28	·80	·48	4·00	1·60
Sats.	·90	·80	·50	·30	2·50	—

Divisors: for Column-Sums, 1·60; for Row-Sums, 2·50.

TABLE XIII. CALCULATION OF SATURATIONS FOR GROUP FACTORS (*P*)

Test	5	6	7	Sums (for checking)
Saturations (Basic Factor) ...	·70	·50	·40	—
Test 5	·53	·43	·32	1·28
Sat. ·70...	·49	·35	·28	1·12
Residuals	·04	·08	·04	·16
Test 6	·43	·41	·28	1·12
Sat. ·50...	·35	·25	·20	·80
Residuals	·08	·16	·08	·32
Test 7	·32	·28	·20	·80
Sat. ·40...	·28	·20	·16	·64
Residuals	·04	·08	·04	·16
Totals of Residuals	·16	·32	·16	·64 = ·80^2
Saturations (Group Factor) ...	·20	·40	·20	·80

Divisor for Totals = $\sqrt{\cdot 64} = \cdot 80$.

(ii) *Determining the Saturations for the Supplementary Group Factors.* The next step is to eliminate the effects of this general or basic factor from the observed correlations. We may begin by reconstructing a theoretical hierarchy for the smallest group from the saturations just obtained for tests 5 to 7, and deduct these theoretical values from the inter-correlations actually observed, as shown in Table XIII. The first set of residuals for these three tests are then factorized by simple summation in the usual way. We can then do the same with the inter-correlations for the larger group, viz., that containing tests 1 to 4. The complete set of factor-saturations thus obtained is shown in Table XI B.

Now our hypothesis is that there are only two factors entering into the inter-correlations for each group of tests. If that hypothesis is correct, then, within each group, the *first* set of residuals (those remaining after the basic factor has been eliminated) should form a submatrix of rank one only (with the smaller group a 3 × 3 submatrix, with the larger a 4 × 4 submatrix). Hence, with an ideal table of correlations like the present the real test for the suitability of our basic factor-saturations is that the *second* set of residuals (those remaining after the appropriate group factor has been eliminated) should vanish identically. With the figures given in Table X this actually happens. In the more general case the fit obtained by using the basic factor given by the first approximation would not be so perfect; further approximations may therefore be required. The routine procedure would consist in subtracting the effects of the group factors alone from the observed inter-correlations for both groups; and then substituting the residuals for the observed inter-correlations in the original table. The entire 7 × 7 table can then be regarded as an approximate hierarchy, and can therefore be factorized by simple summation to obtain improved values for the basic factor. However, this routine procedure may be considerably abridged by making suitable adjustments as they suggest themselves by simple inspection. Alternatively, with submatrices of this size we can write out an algebraic expression for the sum of the residuals in terms of the factor saturations treated as unknowns, set this sum equal to zero, and solve for the unknowns. It is unnecessary to illustrate this procedure here, since it will seldom be needed with any actual table. With an empirical table we cannot expect to find the second set of residuals summing to *exactly* zero. In such a case, therefore, the best procedure would be one that minimized the sum of the squares of the residuals. Again, it is not difficult to construct algebraic equations which will yield a working formula for obtaining the required results by direct calculation. But the calculation would be unnecessarily laborious. For most practical purposes it will be simplest to proceed by successive approximation in the way just suggested, modifying the factor-saturations obtained with the first trial in order to diminish both sets of second residuals so far as possible. Even so, in the vast majority of cases, unless extremely large populations have been tested, the sampling errors will be big enough to render any elaborate adjustments of this kind superfluous.

For a simple example of this procedure applied to an actual correlation table, the reader may refer to the analysis of the Pearson-Macdonell table quoted in the preceding issue of this *Journal* (12, p. 118, Table IX; cf. also Table XV B below). But, as a rule, the matrix of non-overlapping group factors merely serves as an intermediate step towards obtaining the final matrix of overlapping group factors; and consequently in most published articles only the latter is printed.

D. ONE GROUP FACTOR

For the sake of completeness I shall conclude this section by considering how the foregoing principles are to be applied to a correlation table in which there appears to be only a single group factor. Spearman, as we have seen, always doubted the legitimacy of trying to extract more than one group factor from the same table at the same time. When confronted with a table in which the possibility of several group factors called for consideration (as in researches planned, not by himself, but by other investigators), he still preferred to take one group factor at a time, even if this meant discarding a large portion of the table. His method was to select correlations so that these should, so far as possible, relate to two groups of tests only. One group was to be regarded as providing " reference values for *g* " : these, it was assumed, would involve no further factors. The other group would include, or at any rate *might* include, in addition to the general factor, *g*, the supplementary group factor, whose existence he proposed to examine (see above p. 53).

Spearman's procedure, however, sought to investigate simply the presence or absence of this group factor (6, ch. XIII). Even when he admitted its presence, he never went on to estimate its saturation coefficients. Indeed, he devised no formula, and suggested no method, for such determinations.[1]

[1] The procedure which Spearman eventually substituted for group factor analysis is worth describing in some detail. In the current text-books on factorial methods, his criterion is always described, but no explanation is given to show how he actually used it (cf. 10, pp. 68, 112f.; 11, pp. 12f.; Thurstone, *Multiple Factor Analysis*, ch. XII). His own account (6, pp. 223 and xxi) is by no means easy to follow. The fullest explanation of the method is to be found in Hargreaves' *Monograph* on 'The Faculty of Imagination' (*loc. cit. sup.*, 1927, pp. 18–24), where he describes in detail an attempt to isolate a single group factor of 'fluency' from a correlation table including 8 tests of 'intelligence' and 6 tests of 'imagination,' to which tests of Memory, Speed, and Perseveration were added later.

As the investigation was carried out in L.C.C. schools, a preliminary conference was held, as usual, between the supervisor (Spearman), the L.C.C. psychologist (myself), and the head teachers concerned; and the correspondence-file furnishes an instructive account of the alternative suggestions. In discussing the statistical procedure, I had recommended " an extended group factor method " (virtually that described in Section II B above), on the grounds that, not only my own students, but also " Miss Carey, Bradford, Collar, and others had tried it with some success," and that it " dealt most directly with the problem, as envisaged by teachers." Spearman, however, replied that " the true criterion (the tetrad difference criterion) has now been provided " : " even the older [inter-columnar] criterion has shown that the group factors, formerly claimed, have no real existence; and this is now confirmed by the true criterion." " To estimate the size of the supposed group factors merely by accidental departures from a calculated hierarchy is," he maintained, " very unsafe. . . . Every supposed group factor must be scrutinized by itself."

In his *Monograph* Hargreaves accordingly discusses each of the conceivable factors or 'faculties' one by one. He begins by applying the tetrad difference criterion to the tests of Intelligence and Imagination separately; and so shows, or claims to show, that " each of the two groups of tests—Intelligence and Imagination (' Fluency ')—contains a general factor and no second general or group factors." Then, applying the same criterion to the two groups " taken together," he states that " a group factor or second general factor is found." Thus, *one* of the two general factors previously referred to " must be *compound*." Rather summarily, he dismisses the possibility that the Intelligence tests might contain a second factor, " because of evidence that is accumulating against the existence of a group factor in intelligence tests " (he does not specify this evidence, except that on a later page he briefly refers to the possibility of a 'linguistic' group factor in such tests, similar to the 'linguistic ability' I had reported, and rejects it as " most improbable "). Hence in his view the only conclusion remaining is that the initial correlation table contains a *single* group

TABLE XIV. CORRELATIONS BETWEEN PHYSICAL MEASUREMENTS

One Group Factor : Calculation of Basic Factor

Trait		1	2	3	4	Sum	5	6	7	Sum
1. Foot...	...	(·821)	·736	·759	·797	3·113	·363	·339	·206	·908
2. Height	...	·736	(·731)	·661	·800	2·928	·345	·340	·183	·868
3. Finger	...	·759	·661	(·714)	·846	2·980	·321	·301	·150	·772
4. Forearm	...	·797	·800	·846	(·803)	3·246	·289	·305	·135	·729
Sum	...	3·113	2·928	2·980	3·246	12·267	1·318	1·285	·674	3·277
Saturation	...	—	—	—	—	—	·376	·367	·193	·936
5. Face B.	...	·363	·345	·321	·289	1·318	—	·395	·618	—
6. Head L.	...	·339	·340	·301	·305	1·285	·395	—	·402	—
7. Head B.	...	·206	·183	·150	·135	·674	·618	·402	—	—
Sum	...	·908	·868	·772	·729	3·277	—	—	—	—
Total	...	4·021	3·796	3·752	3·975	15·544	—	—	—	—
Saturation	...	·906	·855	·845	·896	3·502	—	—	—	—

Divisor for traits 1–4 = $\frac{15 \cdot 544}{\sqrt{12 \cdot 267}} = 4 \cdot 4381$ Divisor for traits 5–7 = 3·502.

To illustrate the procedure I propose, we may take the Pearson-Macdonell 7 × 7 table, which was analysed in the previous issue of this *Journal* in terms of two group factors superposed upon a single basic factor (see Table XIV). We may imagine a follower of Spearman enquiring whether we cannot after all dispense with one of these two group factors. Can we not assume that there is just a single basic factor determining the general growth of the body and its limbs, and a special group factor confined to the measurements of head and face? Let us therefore see what kind of a fit this hypothesis yields, by calculating the requisite figures according to the principle of simple summation.[2]

We begin by determining the basic saturations for the four body-traits (as we may call them). The first step is to find the sums of the columns of inter- and cross-correlations for these four traits. As usual, the final sums must include the self-correlations[3]: for these, trial values are first inserted, and then corrected in the ordinary way by substituting the squares of the self-correlations so found. The divisor for each column of seven figures is found by dividing the total of all four columns (15·544) by the square root of the sum of the inter-correlations, including the self-correlations (i.e., by $\sqrt{12 \cdot 267} = 3 \cdot 502$). We thus obtain for the saturations 4·021 ÷ 4·4381 = ·906, etc. (see last line of Table XIV).

factor of 'fluency,' entering into the tests of Imagination in addition to the all-pervasive *g*. Since these six tests were shown to obey the tetrad difference criterion, it is inferred that the 'compound general factor' must be what was later called a 'bifid factor' or 'bifactor': i.e., as he says, the two factors must enter into the several tests "in exactly the same proportion." To clinch the conclusion so reached, and determine more precisely the essential nature of the so-called fluency factor, he employs Yule's formula for partial correlation (a method used by Miss Spielman and myself in studying 'creative imagination,' cf. *A Study in Vocational Guidance*, 1926, pp. 23f.).

In the light of his own experience Hargreaves himself eventually admits that he is not altogether content with this procedure. Where single group factors are to be separately disproved, he thinks it satisfactory enough. But "when several factors are likely to be operative, the [tetrad difference] criterion seems to have a limited value: it informs us of the existence of more than one factor, but cannot help to separate the factors out." As to the application of Yule's formula, "apart from the validity of such a proceeding, the mathematical labour involved, when more than one factor has to be eliminated, and when several tests are involved, becomes very great."

---

[2] See Appendix, I, p. 73 below, for the derivation of the formula employed.

[3] Spearman's procedure involves the assumption that the self-correlations are all equal to 1·000: see this *Journal*, I, p. 96. When the cross-correlations include instances of specific overlap, the saturations for the basic factor should in theory be based on the inter-correlations only.

To find the basic factor saturations for the last three traits, we first sum the three 4-figure columns of cross-correlations. For these the divisor is the sum of the four saturations just found (3·502). We thus obtain 1·318 ÷ 3·502 = ·376, etc. The usual checks can be applied: for example, the residuals of each of the four 7-figure columns and of each of the three 4-figure columns should evidently add to zero.

The saturations for the single group factor are obtained by subtracting the theoretical values for the inter-correlations between the three head traits (calculated by multiplying the three saturations just found) from the observed inter-correlations. Since for these traits there are only three observed inter-correlations, and therefore only three residuals, an exact fit can here be obtained. The saturations are computed by the familiar formula $r_{5s} = \sqrt{\frac{r_{56}\, r_{57}}{r_{67}}}$, etc., where r_{56}, r_{57}, r_{67} denote the three residuals: e.g., for trait 5 we have $= \frac{\cdot 257 \times \cdot 545}{\cdot 331} = \sqrt{\cdot 4229} = \cdot 650$, etc.

The factor-saturations so found are shown in Table XV. For comparison I give the figures obtained on the assumption that there are *two* group factors (from 12, p. 118, Table IX). The question now arises: which of the two alternative assumptions furnishes the better fit?

Judged by the absolute amount of discrepancy, the hypothesis of two group factors certainly yields the closest fit. The numerical total of the residuals comes to 0·257 with two group factors, and 0·502—nearly twice as much—with one group factor. Judged according to the principle of least squares, the difference is still more marked; the square-sum in the former case is 0·0045, in the latter 0·0218.

TABLE XV. FACTOR-SATURATIONS

Trait	A. One Group Factor		B. Two Group Factors		
	Size	Head	Size	Body	Head
1. Foot...	·906	—	·719	·501	—
2. Height	·854	—	·688	·464	—
3. Finger	·845	—	·612	·582	—
4. Forearm	·896	—	·578	·816	—
5. Face B.	·376	·650	·508	—	·505
6. Head L.	·367	·395	·495	—	·284
7. Head B.	·193	·839	·260	—	·962
Square-Sum	3·379	1·282	2·270	1·471	1·261
Contribution to Variance (per cent.)	48·3	18·3	32·4	21·0	18·0

But are these differences significant? As the observed correlations were calculated by Pearson's tetrachoric procedure, it is not easy to assign exact figures for the standard errors of the residuals. Macdonell's estimate puts them in the neighbourhood of ± 0·015. Roughly speaking, therefore, any residual that is numerically larger than 0·030 may be regarded as significant. With two group factors no residual reaches this figure. With one group factor seven out of 21 residuals exceed it. The standard deviation with the two-group factor hypothesis is 0·0147 (just about what we should expect by chance); with the one-group factor hypothesis it is more than twice as large, namely, 0·0312.

In point of fact, a preliminary inspection of the table of observed correlations might have led us to anticipate some such result. I have arranged the traits so that the rectangular submatrix of cross-correlations (i.e., the correlations of traits 1–4 with traits 5–7) should conform, so far as possible, to a hierarchical order. It will be noted that they do so with remarkable closeness. If traits 1–4 were also due to the same general factor and nothing more, the inter-correlations for these traits should also be hierarchical, and (what is more important) should exhibit the same order. But this is not the case. The trait which has the *largest* sum for its *inter*-correlations (Forearm) has the *smallest* sum for its *cross*-correlations.

If our critic suggests that perhaps after all we might have been wiser to take, not the three head measurements, but the four body measurements as containing the sole additional group factor, then his suggestion would be still more strongly contradicted by the same argument: for, as simple inspection will show, Head Breadth has the *largest* sum for its inter-correlations, and the *smallest* sum for its cross-correlations. Thus, *both* sets of inter-correlations demand a group factor of their own.

In practice I am inclined to believe that, with correlation tables which are expected to reveal only a single group factor, the investigator will, as a rule, be better advised to start by allowing for two group factors, and then prove that the saturations for one are non-significant. In general, however, I should regard an experiment which yields a correlation table showing one group factor only (or indeed no more than two group factors) as badly planned. Unless the number of persons tested or measured is exceptionally large, and the standard errors in consequence exceptionally small (as they are here), it is usually possible to fit such tables with so many different group factor patterns that the results are bound to be inconclusive.[1]

III. OVERLAPPING GROUP FACTORS

Rotation. When the number of persons tested is large, and especially when the assortment of tests has been chosen to fulfil a practical need rather than to solve a theoretical problem, a pattern of non-overlapping factors is likely to give a somewhat incomplete fit to the correlations observed: in such investigations, as experience shows, we nearly always discover residuals indicative of overlap[2] in the outlying submatrices. But in any case we cannot really assume that an empirically derived factor matrix must always contain a number of saturations that are *exactly* zero. Hence it is desirable to have ready for use some simple arithmetical procedure which will enable us to readjust the schematic pattern of non-overlapping factors so as to yield a closer overall fit.

For this purpose the most obvious device is to rotate the bipolar factor matrix obtained by simple or weighted summation to fit the factor-pattern obtained by the group factor analysis.[3] The use of orthogonal transformations is, of course, a familiar procedure in mathematics; and the idea of arithmetically rotating factors in this way was first suggested by Garnett (5). Graphical methods of rotation have been freely used by Thurstone; but, as Reyburn and others have pointed out, the substitution of graphical for numerical methods is not only less exact, but appears to allow wide room for subjective influences (13).

Example. To illustrate the need for this supplementary procedure and a convenient method of carrying out the calculations let us consider the artificial table of correlations set out on p. 63. It may be regarded as based on a set of 10 cognitive tests applied to a sample of 500 cases (Table XVI).

In the original investigation, planned as part of an experiment in educational and vocational guidance,[4] a series of written, oral, and practical tests was applied to several batches of boys, mostly

[1] It was for this reason that I have not illustrated my arguments in this section by an example drawn from mental measurement. Earlier discussions (e.g., 6, pp. 224–241) provide numerous instances in which one group factor only is discussed. But in nearly every case the probable error is too large to make any clear comparisons profitable. Further, in many cases the wide differences in reliability tend of themselves to introduce a misleading parallelism into the coefficients. For instance, in Hargreaves' table (cited above) the reliability of the tests employed varies from 0·23 to 0·92; and this of itself tends to produce an appearance of consistent proportionality between the inter-correlations and the cross-correlations.

[2] The existence of 'overlap' between group factors was first noted in the analysis of emotional traits, but proved to be a regular feature in analysing correlations between educational abilities (4, pp. 59–60). Spearman uses the term in a somewhat different sense: where I had described two or more tests as involving a *group* factor, he preferred to say that they showed an overlap between their respective *specific* factors (6, p. 82).

[3] Cf. 8, p. 358; 9, p. 308. A more fully worked example is given in *Brit. J. Educ. Psych.*, IX, 1939, pp. 55–65.

[4] The investigation (amongst others) was described in a Report to the Chief Inspector of the L.C.C. Education Department on *Examinations for Junior County Scholarships and Trade Schools*, and I am indebted to the Council for permission to make use of the results.

between the ages of 12½ and 14½, and attending trade schools in or near London. The written tests consisted of two tests of Intelligence, verbal and non-verbal, two papers on 'Arithmetical Problems' and 'Practical Geometry' respectively, and finally two papers on 'Science' requiring brief written answers mostly in essay form—the first, a 'general' paper, dealing mainly with elementary physics and including interpretations of diagrams, the second, dealing more specifically with 'Mechanics' and involving some appreciation of formulæ. These written tests were set to 543 boys, usually as part of a terminal examination at the institute they were attending. Two further tests, Dictation and Mental Arithmetic, were given orally, but required written replies: they were given to the same boys, with the exception of two smaller schools. Two practical tests were included in the scheme—an Assembly test and a test of Practical Science (physics, mechanics, and the like); but these could only be given to a limited number of pupils, amounting in all to 457.

The correlations were calculated for each group separately; and the average correlations (weighted for numbers and reliability) were factorized by the method to be described in a moment. To obtain a more convenient table for purposes of illustration, I have here rounded off the saturation coefficients to a single significant figure; and have then reconstructed an artificial table of 'observed' correlations (Table XVI), which, as a matter of fact, closely resembles the table obtained from the largest group. By taking an example based on an exact set of saturations, we shall have a definite check for the figures ultimately reached by the procedure proposed.

Bipolar Analysis. We begin with a preliminary bipolar analysis. Four factors have been extracted, three of which we may regard as statistically significant. The saturations so obtained are shown in Table XVII A. The interpretation of the three significant factors is quite obvious. The first is a factor of 'general ability'—a weighted average of all the tests; the second divides the tests into non-numerical (tests 1–7) and numerical (tests 8–10); the third divides the non-numerical tests into verbal (tests 1–4) and non-verbal, i.e., practical (tests 5–7). And the factor pattern as a whole suggests that the correlations could probably be fitted by means of a comprehensive basic factor supplemented by three smaller group factors. However, in the lateral submatrices of Table XVI several of the coefficients are much too large for a perfect hierarchical arrangement. Hence we may anticipate some degree of overlapping between the several group factors.

Analysis into Non-Overlapping Group Factors. As before, we may take the bipolar analysis as indicating how the correlation table should be rearranged and where the lines of division between the several groups should fall. The rearrangement has already been carried out in printing the correlations in Table XVI; and the table itself has been partitioned accordingly. The next step is to factorize it in terms of three non-overlapping group factors in order to discover the general frame-work.

An experienced computer would probably realize at once that a satisfactory fit could only be obtained by successive approximation, and would probably start by reducing the more conspicuous enlargements due to the overlapping at the outset. Alternatively he might seek to base the group factor analysis solely on those parts of the lateral submatrices that conform fairly well to hierarchical requirements. The beginner will prefer a simpler and more automatic procedure; and here therefore I shall assume that the method described above (p. 48) is to be applied to the observed correlations taken just as they stand.

The results of factorizing the table by the method described in Section II A are shown in Table XVII B. This second factor-pattern gets rid of negative saturations; but manifestly will not fit the table of observed correlations as well as the bipolar matrix. It is easy to see, for example, without actually reconstructing a theoretical table of correlations from the factor-saturations, that the residuals for tests 2 and 3 with 10 and for tests 2 and 9 with 6 will be fully significant. What we now require therefore is a readjusted group factor matrix derived from the previous bipolar matrix (Table XVII A), but conforming to the schematic pattern exhibited by Table XVII B.

We could, of course, use these two tables to calculate a transformation matrix: if this transformation matrix proved to be strictly orthogonal, then the fit obtained by the rotated factors would be as good as that obtained direct from Table XVII A itself, and the new factors would themselves be orthogonal. However, the correlation tables reconstructed from the group and bipolar factors (sections A and B) would be by no means identical; and consequently the transformation matrix would be far from orthogonal. We might go on to modify the transformation

TABLE XVI. OBSERVED CORRELATIONS

Test	1	2	3	4	Sum	5	6	7	Sum	8	9	10	Sum
1. Verb. Int.	(·82)	·76	·48	·29	—	·63	·44	·37	1·44	·72	·54	·21	1·47
2. Sci. 1 (Gen.) ...	·76	(·84)	·52	·32	—	·60	·44	·40	1·44	·64	·52	·28	1·44
3. Sci. 2 (Mech.)...	·48	·52	(·35)	·21	—	·35	·22	·23	·80	·44	·35	·25	1·04
4. Dict.	·29	·32	·21	(·13)	—	·21	·13	·14	·48	·24	·18	·12	·54
Sum	2·35	2·44	1·56	·95	7·30	1·79	1·23	1·14	4·16	2·04	1·59	·86	4·49
5. Non-Verb. Int.	·63	·60	·35	·21	1·79	(·53)	·43	·32	—	·56	·46	·14	1·16
6. Mech. Ass. ...	·44	·44	·22	·13	1·23	·43	(·42)	·27	—	·40	·38	·07	·85
7. Sci. 3 (Pract.) ...	·37	·40	·23	·14	1·14	·32	·27	(·21)	—	·32	·28	·11	·71
Sum	1·44	1·44	·80	·48	4·16	1·28	1·12	·80	3·20	1·28	1·12	·32	2·72
8. Prob. Arith. ...	·72	·64	·44	·24	2·04	·56	·40	·32	1·28	(·80)	·68	·40	—
9. Prac. Geom. ...	·54	·52	·35	·18	1·59	·46	·38	·28	1·12	·68	(·65)	·42	—
10. Arith. Oral ...	·21	·28	·25	·12	·86	·14	·07	·11	·32	·40	·42	(·49)	—
Sum	1·47	1·44	1·04	·54	4·49	1·16	·85	·71	2·72	—	—	—	—
Col. Total	2·91	2·88	1·84	1·02	8·65	2·95	2·08	1·85	6·88	3·32	2·71	1·18	7·21
Divisors	3·301					4·333				4·208			
Saturations ...	·882	·872	·557	·309	2·620	·681	·480	·427	1·588	·789	·644	·280	1·713

matrix, so as to make it orthogonal. But this procedure would, as a rule, be rather laborious; and in any case is suited only to more experienced computers.[1]

Calculation of Rotation Matrix from Condensed Factor Matrices. If, however, instead of taking the initial table of observed correlations, we proceed to reconstruct a new correlation table from Table XVII B as it stands, and if we then subject this reconstructed table to a bipolar analysis, we should have two factor matrices, bipolar and group, which furnished *exactly* the same set of correlations. The rotation matrix for transforming the first factor matrix into the second would then be perfectly orthogonal. Since, however, we are concerned with factor matrices containing 4 columns only, the order of the rotation matrix will be 4 × 4; and consequently, to calculate such a rotation matrix, we do not need group- and bipolar-matrices containing 10 rows; 4 will be sufficient. Let us therefore condense the group factor matrix into a 4 × 4 matrix; from this let us reconstruct a symmetrical product-sum matrix, by the procedure regularly used for reconstructing a theoretical correlation table. The 4 × 4 matrix so obtained will then, as a rule, be identical with the table we should get if we condensed the reconstructed 10 × 10 correlation table to a 4 × 4 table of summed correlations.

Accordingly, with the present investigation we can begin by splitting the first group of 4 tests into two subgroups, the first containing test 1 only and the second tests 2, 3, and 4. With the other three groups we can take the sums as they stand. We shall then work with the sums of the factor-saturations for these four groups (see Tables XVII B and XVIII A). The 4 × 4 matrix of summed correlations constructed from these sums is shown in Table XVIII B. This is now subjected to a bipolar analysis; and the saturations so obtained are given in Table XVIII C. These should be identical with the sums of the bipolar factor-saturations we should have extracted from the complete 10 × 10 correlation matrix.[2]

It is now an easy matter to calculate the transformation matrix, required for rotating the figures in Table XVIII A to those in Table XVIII C. There is no need to determine an inverse. The fact that the sums of the non-overlapping group factors include a number of zeros renders a direct calculation, by the ordinary method of solving linear equations, perfectly simple (for the method of calculation see Appendix, p. 75).

The transformation matrix thus obtained is given in Table XIX A. The reader can easily satisfy himself that it is orthogonal.

Rotation of Original Factors. (*a*) *First Approximation.* We now use this transformation matrix to postmultiply the *original* bipolar-matrix (Table XVII A). This yields the set of overlapping group factors as shown in Table XX A. The pattern of saturations is fairly satisfactory; and, if the number tested had been small, it could be accepted as it stands.

It reveals several clear instances of positive overlap, e.g., tests 7 and 10 in the second factor (*V*), test 9 in the third factor (*P*). But it also yields three negative saturations over 0·100 (for tests 3, 8, and 10). Such negative values are apt to occur in sections where there is an appreciable positive

---

[1] It has, as a matter of fact, been used in several research-theses (e.g., Miss Stevenson's analysis of tests applied in the Navy during the war). The orthogonalizing process is best carried out as follows. The general principle will be to restrict the largest adjustments to those rows or figures of the transformation which will multiply the smallest columns in the bipolar matrix. To begin with, we should leave unchanged the first row in the transformation matrix which multiplies the first column of the bipolar matrix; we should alter the least important figure in the second row, so that the product-sum of the first and second rows is zero; we should then modify the two least important figures in the third row so that its product-sum with the first two rows is zero; and so on. Finally, for the last and least important row of all we substitute figures which will make the sums of the squares in each column equal to unity.

[2] As in the 'sum method' of factor analysis (this *Journal*, II, pp. 62f.). Certain cautions have to be exercised in regard to the reversal of signs. The beginner is advised in his first efforts to factorize the complete 10 × 10 artificial correlation table as a check: with certain modes of pooling and with the ordinary rules for sign reversal, he may find that the totals for later factors differ slightly when obtained from the full table and from the summed table respectively. As a rule, such minor divergences will not appreciably affect the final results; and in any case, those results will generally be treated as a first approximation. Hence the second approximation can be trusted to eliminate any defects due to the provisional mode of pooling.

TABLE XVII. FACTOR-SATURATIONS FOR BIPOLAR AND NON-OVERLAPPING GROUP FACTORS

Test	A. Bipolar Factors				Square-Sum	B. Group Factors				Square-Sum
	I	II	III	IV		*G*	*V*	*P*	*N*	
1. Verb. Int.	·8513	·1968	·0085	·2378	·8200	·8816	·0579	—	—	·7806
Sum	·8513	·1968	·0085	·2378	—	·8816	·0579	—	—	—
2. Sci. 1 (Gen.)	·8610	·2442	·1737	−·0945	·8400	·8725	·2381	—	—	·8180
3. Sci. 2 (Mech.) ...	·5503	·0405	·2134	·0075	·3500	·5574	·1734	—	—	·3480
4. Dict.	·3188	·0839	·1460	·0036	·1300	·3090	·2184	—	—	·1432
Sum	1·7301	·3686	·5331	−·0834	—	1·7389	·6299	—	—	—
5. Non-Verb. Int. ...	·6846	·1796	−·1612	·0558	·5300	·6807	—	·2420	—	·5220
6. Mech. Ass.	·5179	·1625	−·3309	−·1262	·4200	·4800	—	·4344	—	·4191
7. Sci. 3 (Pract.) ...	·4289	·1287	−·0495	−·0840	·2100	·4267	—	·1480	—	·2040
Sum	1·6314	·4708	−·5416	−·1544	—	1·5875	—	·8244	—	—
8. Prob. Arith.	·8416	−·2174	−·0738	·1978	·8000	·7890	—	—	·3712	·7603
9. Prac. Geom.	·7218	−·3192	−·1589	−·0433	·6500	·6440	—	—	·4615	·6277
10. Arith. Oral	·4030	−·4996	·2328	−·1545	·4900	·2804	—	—	·5901	·4268
Sum	1·9664	−1·0362	·0000	·0000	—	1·7134	—	—	1·4228	—
Square-Sum	4·1827	·5810	·3196	·1569	5·2400	3·9364	·1379	·2692	·6990	5·0425
Contribution to Variance (per cent.)	41·8	5·8	3·2	1·6	52·4	39·4	1·4	2·7	7·0	50·4

overlap in the final saturations, and where the provisional non-overlapping matrix gives zero saturations: the negative saturations are then arithmetically necessary to balance the positive saturation that represents the overlap. By taking the correlations in the lateral submatrices just as they stand, we have in effect treated the augmented coefficients as non-significant results of sampling fluctuation. That means that our saturations represent averages, and the residuals represent non-significant deviations, positive and negative, about such averages. At this point, therefore, it is desirable to make a rough test of the significance of the larger positive and negative saturations which take the place of the previous zero saturations, and, if they approach the size of the probable error, proceed to a further approximation.

(*b*) *Second Approximation.* There are various ways in which an improved rotation can be effected. Two may be described as most convenient for general purposes.

(i) The rotated matrix just obtained reveals where the overlaps occur and what is their probable size. We can therefore go back to the original correlation matrix, and deduct from the augmented coefficients the amount apparently due to the overlap. The analysis into non-overlapping group factors can then be repeated with the smoothed correlation table; and the whole of the subsequent calculations carried out afresh. This is probably the simplest routine for the beginner.[1]

(ii) An alternative method is to modify two or three figures in the rotation matrix, so as to suppress the larger negative saturations, and then orthogonalize the columns afresh. Thus the negative saturations in the second and the third factors are obviously due to the large negative figures in the second and third columns of the rotation matrix (− ·8379 and − ·8598): these can very readily be reduced. Minor adjustments must then be made in other figures to ensure that the product sums of the different pairs of columns are exactly zero; and each of the readjusted columns must then be normalized to ensure that the sum of its squares is 1·000.

Final Results. An adjusted rotation matrix obtained in this way[2] is shown in Table XIX B. This is used, as before, to postmultiply the original bipolar-matrix (Table XVII A). The resulting factor saturations are shown in Table XX B. For comparison I append in the last four columns the factor-saturations from which the original correlation matrix was actually constructed. It will be observed that the agreement is remarkably close. With three minor exceptions the discrepancies between the calculated values and the true values are all less than ± 0·015; and, of course, with a further series of approximations we could get as close as we desire.

The interpretation of the factors is fairly obvious: the first is a 'basic' factor of 'general ability'; the rest represent 'special abilities' for verbal, practical, and numerical work. The instances of overlap are perfectly intelligible, when we recall the nature of the tests.

Constancy of Group Factor Matrices. In any field of enquiry where the investigators may wish to assign a concrete meaning to each of the factors extracted, it is desirable to determine the factors in such a way that their interpretation shall remain the same, even when they are calculated afresh from a modified battery, to which certain tests have been added, or from which certain tests have been omitted. In previous accounts of the group factor method, I have claimed that it will fulfil this

---

[1] As noted above, the more experienced computer would probably have observed at the very outset which of the coefficients disturbing the hierarchical arrangement showed significant augmentations, and would have made tentative deductions before undertaking the first group factor analysis. Even with a brief inspection of Table XVI, it was fairly obvious that test 10, for example, shows exaggerated correlations with tests 2, 3, and 4. When these coefficients have been reduced, several smaller augmentations are readily detected (e.g., tests 2 and 8 with tests 5, 6, and 7).

[2] This and the remaining calculations were made for me by a former research student, Mr. F. R. Llewellyn, who knew nothing of the true values of the saturation coefficients or of how the original correlation table had been constructed. I felt that, were I myself to make the requisite adjustments, I might be consciously or unconsciously influenced by my previous knowledge of the true values —or at least that it might be supposed that some such influence had operated. Having reached the figures shown in Table XVII B, Mr. Llewellyn began to suspect that he was working on an artificial problem. In the hope of suppressing the negative saturation for test 6, he spontaneously carried out a further approximation. Working with six decimal figures he then obtained the true values to within five decimal places (including the negative saturation).

TABLE XVIII. FACTOR ANALYSIS OF PRODUCT-SUMS

Tests	A. Summed Group Factor-Saturations (G_s)				B. Product-Sums ($R = G_s G_s'$)				C. Bipolar Analysis of R (B_s)			
	G	V	P	N	1	2–4	5–7	8–10	I	II	III	IV
1	·8816	·0579	·0000	·0000	3·1644	1·6358	2·7846	3·0054	·8506	·1568	·1062	·1453
2–4	1·7389	·6299	·0000	·0000	1·6358	·9042	1·3755	1·4846	1·7352	·4177	·4628	−·1453
5–7	1·5875	·0000	·8244	·0000	2·7846	1·3755	3·1996	2·7200	1·6300	·4680	−·5690	·0000
8–10	1·7134	·0000	·0000	1·4228	3·0054	1·4846	2·7200	4·9601	1·9681	−1·0425	·0000	·0000

TABLE XIX. ROTATION MATRICES

Factors	A. First Approximation ($T_s = G_s^{-1} B_s$)				B. Second Approximation			
	G	V	P	N	G	V	P	N
I	·9575	·1113	·1334	·2303	·9296	·2071	·1769	·2482
II	·1641	·2100	·2517	−·9304	·1917	·1105	·2013	−·9542
III	·0882	·4914	−·8598	−·1063	−·0778	·8415	−·5337	−·0308
IV	·2199	−·8379	−·4236	−·2648	·3051	−·4865	−·8020	−·1645

TABLE XX. FACTOR-SATURATIONS FOR OVERLAPPING GROUP FACTORS

Test	A. First Approximation					B. Second Approximation					C. True Values				
	G	*V*	*P*	*N*	Square-Sum	*G*	*V*	*P*	*N*	Square-Sum	*G*	*V*	*P*	*N*	Square-Sum
1. Verb. Int. ...	·901	−·056	·053	−·049	·820	·901	·086	−·010	−·014	·820	·900	·100	—	—	·820
2. Sci. 1 (Gen.)...	·859	·312	·068	−·019	·840	·805	·398	·187	−·009	·840	·800	·400	·200	—	·840
3. Sci. 2 (Mech.)	·554	·169	−·103	·067	·350	·505	·294	−·014	·090	·350	·500	·300	—	·100	·350
4. Dict.	·331	·122	−·063	−·020	·130	·301	·195	−·008	−·059	·130	·300	·200	—	—	·130
5. Non-Verb. Int.	·683	−·013	·251	−·008	·530	·700	−·001	·198	−·006	·530	·700	—	·200	—	·530
6. Mech. Ass. ...	·466	·033	·448	·035	·420	·499	−·095	·402	·004	·420	·500	−·100	·400	—	·420
7. Sci. 3 (Pract.)...	·409	·120	·168	·007	·210	·402	·102	·195	−·001	·210	·400	·100	·200	—	·210
8. Prob. Arith. ...	·807	−·153	·036	·353	·800	·807	·009	−·014	·386	·800	·800	—	—	·400	·800
9. Prac. Geom....	·614	−·029	·171	·492	·650	·609	−·008	·183	·496	·650	·600	—	·200	·500	·650
10. Arith. Oral ...	·289	·185	−·205	·576	·490	·212	·300	−·030	·595	·490	·200	·300	—	·600	·490
Square-Sum ...	3·930	·216	·388	·706	5·240	3·773	·398	·309	·760	5·240	3·730	·410	·328	·780	5·240
Contrib. to Variance (per cent.)	39·3	2·2	3·9	7·1	52·4	37·7	4·0	3·1	7·6	52·4	37·3	4·1	3·3	7·8	52·4

requirement within the limits of the probable error. May I therefore conclude my description by a brief illustration of this point?

‘ Let us take the last correlation table (Table XVI), and omit the final group of tests, i.e., the ‘numerical’ tests, 8, 9, and 10. The remaining 7 × 7 table can now be factorized by the method described in Section C for the case of two group factors only. The saturations obtained first by preliminary bipolar analysis and then the final analysis in terms of non-overlapping group factors are set out in Table XXI B.

TABLE XXI. FACTOR ANALYSIS OF CURTAILED BATTERY

Tests	A. Bipolar Factors			B. Group Factors		
	I	II	III	*G*	*V*	*P*
1. Verb. Int.	·874	·076	·225	·881	·121	—
2. Sci. 1 (Gen.) ...	·895	·138	−·144	·881	·219	—
3. Sci. 2 (Mech.) ...	·542	·211	−·042	·489	·345	—
4. Dict.	·330	·140	−·039	·294	·223	—
5. Non-Verb. Int. ...	·708	−·139	·098	·704	—	·178
6. Mech. Ass.	·542	−·354	−·029	·484	—	·451
7. Sci. 3 (Pract.) ...	·447	−·072	−·069	·448	—	·093

On comparing the figures reached with those given in Table XX B, it is obvious that the agreement is exceedingly close. In only one instance (test 2) does the discrepancy exceed 0·100. We could, if we wished, attempt a further approximation; or, if the small residuals were statistically significant, seek a slightly closer fit by a rotation which would substitute small saturations for the zeros.

The initial correlation table, as we know, includes several instances of overlapping. It is this overlapping that has caused the small discrepancies. With only two group factors we can scarcely expect to detect or to allow for it with any degree of accuracy. The ideal procedure would be, as before, to invoke the principle of least squares. The calculations so entailed would be quite practicable, but exceedingly lengthy; and in general the results would differ but little from those achieved by the simpler method described above.

With artificial tables, containing no overlapping whatever, the saturations reached by the group factor analysis would necessarily be identical throughout, no matter how many groups of tests we added or removed (provided, of course, at least two groups remained), and no matter what tests we dropped from the several groups (provided, of course, at least two tests were left in each). This will be obvious if the reader turns back to Section II C, and considers the analysis of Table X, which in point of fact was constructed from the group factor saturations shown in Table XX C with the overlapping eliminated (cf. Table XI B). In such cases therefore the figures would remain perfectly constant or invariant.

Simple Structure. Thurstone's conception of a unique ‘simple structure’ is in many ways closely similar to what some of us in earlier publications had described as a ‘group factor pattern.’ A ‘simple structure’ should, we are told, satisfy three conditions: if V denotes the factor matrix of r columns, then “(1) each row of V should have at least one zero [saturation]; (2) each column of V should have at least r zeros; (3) for every pair of columns there should be at least r traits whose entries vanish in one column but not the other.”[1] It will be observed that the second of these conditions precludes any ‘general’ or ‘basic’ factor, since by definition such a factor must have a non-zero saturation for every trait in its column.

In addition there are bound to be further differences in the distribution and saturations of the other factors, resulting partly from the differences in the conditions and partly from differences in method. The ‘primary factors’ making up a ‘simple structure’ are derived from a preliminary ‘centroid’ analysis by a series of ‘rotations’ (not necessarily orthogonal). The ‘rotations’ are carried out, not arithmetically, but by a graphical procedure which is usually lengthy and inevitably somewhat rough; nor is any use made of the sign-pattern of the centroid matrix to determine the essential limits of the several group factors. Their limits are in effect largely dependent on the choice of the computer at each stage, as he selects this or that pair or triplet of variables for the next rotation.

[1] *Vectors of Mind* (1935), p. 156.

Both the similarities and the differences can readily be observed in the concrete examples where both methods of analysis have been applied to the same correlation table, e.g., where Professor Thurstone has re-analysed a set of traits, previously factorized by the group factor method, and where a set of traits previously factorized in terms of Thurstone's simple structure has been re-analysed by the group factor method.[1] Wherever a strict orthogonal rotation would inevitably yield a general factor, Professor Thurstone prefers to analyse the correlation matrix in terms of an 'oblique' simple structure, i.e., in terms of correlated factors; and then (as he has suggested in his more recent contributions) a further analysis can be undertaken, if we wish, to extract a set of 'second order' factors, one of which will be the general factor.

Thus with Thurstone's procedure three consecutive analyses are required to do what can frequently be done by a single analysis, if we seek a simple group factor pattern.[2] In more complex cases, where there is considerable overlapping, there seems no reason to insist on conditions 1 and 3. Thus, instead of assigning the zero (or non-significant) saturations solely to satisfy certain *a priori* requirements, expressed arbitrarily for all cases in terms of the empirical number of factors (a number which is largely an accidental result of the size of the sample), my own method is to decide their location by the aid of the preliminary bipolar pattern. That pattern, of course, is in turn really determined by the explicit or implicit hypothesis, which has guided the investigator in choosing his particular sets of tests or traits at the very outset of his experiment (provided that his hypothesis is applicable to the data so procured).

There are other less obvious differences depending partly on our mathematical and partly on our psychological preconceptions. But these would require discussion at some length, and must therefore be deferred to a later paper. Here I have said enough to show that there are clear differences both in method and in results. May I add that I do not wish to deny that there may be problems for which Thurstone's procedure is the best, just as there may be other problems for which the bifactor method may be more appropriate?

In conclusion let me repeat that this article has been primarily concerned with practical working methods only. There are a number of theoretical issues (e.g., the precise relation between the bipolar classification and the group factor classification, which is by no means simple in the case of the smaller bipolar factors, the criteria for choosing between a bipolar and a group factor hypothesis, the method of planning experiments for group factor analysis, etc.) which I have no space to touch on here.

Advantages of the Group Factor Method. In my previous discussions of various factorial procedures I have suggested that the factors extracted should, if possible, fulfil three main conditions: paucity, unequivocality, and stability. From our survey of the group factor method, it may, I think, be fairly claimed that its results will in general satisfy these requirements. First, the values given for the group factor-saturations will be (within the limits imposed by the sampling error) either positive or zero; hence each of the traits or tests to be factorized will be accounted for in terms of the minimum number of uncorrelated factors. Secondly, each factor will be (within the limits of sampling) uniquely determined. Thirdly, the factors will remain stable, if not absolutely invariant, even when the battery of tests or traits is modified, e.g., when a comparatively small battery is enlarged by the addition of more tests or more groups of tests, or when a large battery is curtailed by the omission of tests. Occasionally, it is true, the results obtained may fail to preserve all three conditions. But such cases are, I believe, exceptional. They occur, in my experience, only where the investigator has selected his battery unwisely, and where, as a consequence, *no* factorial procedure could really be expected to yield satisfactory results.

---

[1] In *Psychometrika*, XI (1946), p. 20, Professor Thurstone presents an alternative analysis for a table of physical measurements. In *Brit. J. Educ. Psych.*, IX, 1939, pp. 270–276, Dr. Eysenck has carried out a group factor analysis of the correlation table given by Professor Thurstone in his work on *Primary Abilities*.

[2] In the second edition of his book (11, p. 297) Professor Thomson gives a 'second order analysis' of a 7-variable version of Thurstone's box problem, and shows how it leads to a "general pattern containing a 'general factor *plus* an orthogonal simple structure.'" But this is simply the old basic-plus-group-factor pattern, reached by a more circuitous route. Reached by the group factor method the pattern of saturations is very much the same, but the fit is closer (cf. Burt, C., *Brit. J. Educ. Psych.*, table on p. 47).

IV. SUMMARY AND CONCLUSIONS

1. Many tables of correlations, covariances, or product-sums, particularly in psychology, can be expressed in terms of a simple and intelligible factor-pattern, if the observed coefficients are treated as the result of (i) a small number of positive 'group' factors each entering into a few of the variables only, superposed upon (ii) a single 'basic' factor entering into all the variables. The aim of this article has been to make a comprehensive and systematic survey of the working methods that have proved to be most effective for this purpose.

2. By definition the significant saturations for both the 'basic' and the 'group' factors will have positive values only. Two modes of approach are available for securing this result—an addition method and a multiplication method. Reasons are given for preferring the former. With that method the essential principle is to condense the matrix of observed correlations by simple summation. This yields a direct mode of analysis which is both easy and speedy, and in the simplest cases provides an adequate set of factors after a single analysis.

3. Where there is any doubt about the grouping of the tests, it is advisable to begin with a preliminary bipolar analysis. The pattern of signs so obtained supplies an objective indication for the appropriate partitioning of the table of observed correlations.

4. Slightly different working procedures are described for the cases of one, two, three, and more than three group factors. The simplest and commonest case is that of three group factors. The hypothesis of subdivided factors, however, will generally require the determination of paired group factors; and with larger tables there is no fixed limit to the number of factors that can be determined.

5. Where the sample is small, it is usually sufficient to seek non-overlapping group factors. Where the sample is large, it will generally be necessary to allow for overlapping. This can best be done by an arithmetical rotation, which appears less liable to subjective influences than graphical rotation. A suitable rotation matrix can be quickly computed by condensing the group factor matrix by summation, and then making a bipolar analysis of the product-moment table reconstructed from these sums.

6. Finally, it is argued that the factors so found are uniquely determined, economical in number, and virtually invariant when the battery is enlarged or diminished.

REFERENCES

1. Burt, C. (1909). 'Experimental Tests of General Intelligence.' *Brit. J. Psych.*, III, 94–177.
2. Simpson, B. R. (1912). 'Correlations of Mental Abilities.' *Columbia University Contributions to Education*, No. 53, Teachers' College.
3. Bickersteth, M. E., and Burt, C. (1916). 'An Analysis of Results of Mental and Scholastic Tests.' *Report Conf. Educ. Assoc.*, 30–37.
4. Burt, C. (1917). *The Distribution and Relations of Educational Abilities.* L.C.C. Report No. 1868. London: P. S. King & Son.
5. Garnett, J. C. M. (1919). 'General Ability, Cleverness, and Purpose.' *Brit. J. Psych.*, IX, 345–366.
6. Spearman, C. (1926). *Abilities of Man.* London: Macmillan.
7. Burt, C. (1936). 'The Analysis of Examination Marks,' ap. Hartog, P., Rhodes, E. C., and Burt, C. *The Marks of Examiners.* London: Methuen.
8. Burt, C. (1938). 'Factor Analysis by Submatrices.' *J. Psych.*, VI, 339–375.
9. Burt, C. (1940). *Factors of the Mind.* London: University of London Press.
10. Holzinger, K. J., and Harman, H. H. (1941). *Factor Analysis.* Chicago: University of Chicago Press.
11. Thomson, G. H. (1946). *The Factorial Analysis of Human Ability* (2nd ed.). London: University of London Press.
12. Burt, C. (1949). 'Alternative Methods of Factor Analysis.' *Brit. J. Psych.* (*Stat. Sect.*), II, 98–121.
13. Reyburn, H. A., and Raath, M. J. (1949). 'Simple Structure: A Critical Examination.' *Brit. J. Psych.* (*Stat. Sect.*), II, 125–133.

APPENDIX

I. *Basic Formulæ*

Postulates and Definitions. Let x_i and y_i be the observed measurements of any given individual, i, for any pair of correlated traits, e.g., his marks as obtained from two mental tests. We shall assume that such measurements can always be expressed as linear functions (weighted sums) of a set of uncorrelated factors. Accordingly, let us write

$$x_i = w_{xb}b_i + w_{xc}c_i + \dots + w_{xs}s_i \,, \quad . \quad . \quad . \quad . \quad . \qquad \text{(i)}$$

$$y_i = w_{yb}b_i + w_{yc}c_i + \dots + w_{zt}t_i \,, \quad . \quad . \quad . \quad . \quad . \qquad \text{(ii)}$$

where b_i, c_i, s_i, and t_i are i's hypothetical measurements for the several factors and remain constant for the same individual in all the different tests, and where w_{xb}, w_{xc}, w_{yb}, w_{yc}, etc., are the weights or 'factor-loadings' for the several tests, and remain constant for the same test in all the different individuals, i.e., they denote the proportionate amounts of the factors b, c, etc., in the tests x and y.

Evidently the weights may take positive, negative, or zero values. A factor which has positive weights for some of the tests and negative for the remainder will be called a *bipolar* factor. A factor which has positive weights for some of the tests and zero values for the remainder will be called a *group* factor. A factor which has positive weights for every test will be called a *basic* factor if all (or practically all) the weights for the other factors are positive, but smaller: otherwise (e.g., if there are bipolar factors) it will be called a *general* factor.[1] A factor which has zero weights for all except one particular test will be called a *specific* or 'unique' factor: when the tests are not repeated, it will include the error factor.

The measurements will be rendered comparable, and the ensuing formulæ greatly simplified, if we suppose the x, y, b, c, ... s, and t are all expressed in unitary standard measure: so that $\Sigma x^2 = \Sigma y^2 = \dots = 1$; and therefore $w^2_{xb} + w^2_{xc} + \dots + w^2_{xs} = 1$; and similarly for the weights in y.

With these assumptions the following theorems can readily be deduced.

(i) *Saturation Coefficients.* On substituting for x from equation (i), we have—

$$r_{xb} = \frac{\Sigma xb}{\sqrt{\Sigma x^2}\,\sqrt{\Sigma b^2}} = \Sigma\,(w_{xb}b + w_{xc}c + \dots + w_{xs}s)b = w_{xb} \,, \quad . \quad . \qquad \text{(iii)}$$

since $\Sigma\, cb$, ... $\Sigma\, sb$, will be zero (because by hypothesis the factors are uncorrelated), and $\Sigma\, b^2 = 1$. Similarly, $r_{xc} = w_{xc}$; $r_{yb} = w_{yb}$; $r_{yc} = w_{wc}$. Thus, the correlation between a given test and any one of its component factors turns out to be identical with the weight assigned to that factor in the theoretical construction of the test. (Spearman's statement that the weight is proportional to the *square* of the correlation between the test and the factor springs from a confusion of this case with that of a correlation between two tests where the weight of the common factor is assumed to be the same for both tests.)

(ii) *Product-Theorem.* Similarly, on substituting for x and y from equations (i) and (ii), we have

$$r_{xy} = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} = w_{xb}w_{yb} + w_{xc}w_{yc} + \dots = r_{xb}r_{yb} + r_{xc}r_{yc} + \dots \quad . \qquad \text{(iv)}$$

(iii) *Bipolar Analysis.* Now let us suppose we have n tests dependent on the general or basic factor, b, and on one supplementary factor, c (say), which we will first suppose to be common to *all* the n tests. Then, taking the total of the xth column of correlations and the grand total of all the correlations, we have (from (iv))

$$\frac{\Sigma\, r_{xk}}{\sqrt{\Sigma\Sigma\, r_{kk'}}} = \frac{r_{xb}\Sigma\, r_{kb} + r_{xc}\Sigma\, r_{kc}}{\sqrt{(\Sigma\, r_{kb})^2 + (\Sigma\, r_{kc})^2}} = r_{xb}\,, \quad (k, k' = 1, 2, \dots x, \dots n) \quad . \quad . \qquad \text{(v)}$$

provided $\Sigma\, r_{kc} = 0$. Thus, for equation (v) to hold, c must be a bipolar factor, and the sum of its positive and negative saturations must be zero. This treats the saturation coefficient, so calculated, as a kind of average. Equation (v) gives the ordinary formula for factor analysis by 'simple summation.' It can obviously be generalized to include the case where there is more than one supplementary factor, like c.

(iv) *Group Factor Analysis.* The working methods that follow depend on what I call the 'principle of condensation' (see above, p. 45). The factor matrix and the correlation matrix are first partitioned conformably, according to the grouping of the traits. The column-sections of the factor matrix, and the submatrices of the correlation matrix, are then summed. The matrices of sums so obtained I call 'condensed' matrices. It can then be shown that the condensed correlation matrix (i) will have the same rank as the original table, and (ii) can be deduced from the condensed factor matrix, i.e., from the summed saturations. For, let R_{12} be any rectangular submatrix obtained by partitioning the correlation table; let $R_{12} = F_1F_2'$, where F_1 and F_2 are the corresponding sections of the factor matrix; and let w_0' denote the summation operator $[1, 1, \ldots 1]$. Then $w_0'R_{12}w_0 = w_0'F_1F_2'w_0$, where $w_0'F_1$ and $w_0'F_2$ are the sums of the column-sections.[1] Further, if W' denotes the semi-condensation operator (whose rows consist of 1's or 0's), then $W_0'RW_0 = W_0'FF'W_0$.

---

[1] There is no *a priori* reason why bipolar and group factors should not appear in the same factor matrix. All that is necessary is that the extraction of each factor should reduce the rank of the matrix by one: and this, as is easily shown, is a very broad requirement. The methods here described can readily be combined.

(*a*) *One Group Factor.* Let us now suppose that c is common only to a limited group of similar tests. And for simplicity, let us take the case of four tests only, with the first two tests alone involving the group factor c. Expressing the inter- and cross-correlations in terms of the factor-saturations, the whole correlation table, and the sums which result from its condensation, may be written as in Table XXII. Here the capital letters denote sums of factor-saturations, i.e.,

$$B_1 = \sum_1^2 r_{kb}\,,\ B_3 = \sum_3^4 r_{kb}\,,\ C_1 = \sum_1^2 r_{kc}\,.$$

To find the *basic* factor-saturations from the column totals ($r_{1b}B_3$, $r_{3b}(B_1 + B_3)$ . . ., etc.), we must first find the divisors B_3 and $(B_1 + B_3)$. We can obtain B_3 by taking the square root of the sum for the South-East quadrant, viz., $\sqrt{B_3^2}$; and we can obtain $(B_1 + B_3)$ by dividing the grand total of the last two columns by B_3.

To find the factor-saturations for the *group* factor c, we must subtract the 'hierarchy,' formed by multiplying the basic factor-saturations, r_{1b} and r_{2b}, from the North-West quadrant. This leaves a second 'hierarchy,' dependent solely on the group factor c, which can be factorized in the usual way.

(*b*) *Two Group Factors.* Let us now suppose that the South-East quadrant of Table XXII is modified to include a group factor of its own, whose saturation-sum is D_3. The divisors for the row- and column-sums of the North-East quadrant will be B_1 and B_3. The simplest procedure is to determine the ratio B_1/B_3 provisionally (e.g., from the bipolar analysis), calculate C_1 and D_3, and then, if necessary, find improved values for B_1 and B_3 by successive approximation.

(*c*) *Three Group Factors.* Let us take B_1, B_2, B_3 to represent the sums of the basic factor-saturations within the three groups. Then the sums of the non-diagonal submatrices will be B_1B_2, B_1B_3, and B_2B_3. To find B_3, for example, we now calculate $\sqrt{\dfrac{B_1B_3 \,.\, B_2B_3}{B_1B_2}} = \sqrt{B_3^2}$. The divisor for the first column will then be $B_2 + B_3 = \sqrt{\dfrac{B_1B_2 \,.\, B_2B_3}{B_1B_3}} + \sqrt{\dfrac{B_1B_3 \,.\, B_2B_3}{B_1B_2}} = \sqrt{B_2B_3}\left\{\sqrt{\dfrac{B_1B_2}{B_1B_3}} + \sqrt{\dfrac{B_1B_3}{B_1B_2}}\right\}$, which is the formula used on p. 48 above. Similarly for the other divisors.

(*d*) *More than Three Group Factors.* For the general case we simply use the analogue to equation (v), taking the *sums* of the coefficients in the submatrices (obtained by condensation) instead of the individual coefficients themselves.

Method of Least Squares. The foregoing proofs assume that the observed correlations are the *exact* resultants of the basic and group factors. With an empirical table this will not be strictly true. For any given test, however, we can plausibly assume that the *averages* of its observed and of its true correlations are virtually the same. For example, taking test 3 in Table XXII below, we can assume that the average correlation $\bar{r}_{3k}'$ (where the dash denotes an *observed* value) corresponds sufficiently closely with the average of the 'true' correlations ($\bar{r}_{3k}$). Now, with the 'true' values $n\,\bar{r}_{3k} = \Sigma r_{3k} = r_{3b}(B_1 + B_3)$ exactly. If, therefore, we assume $\bar{r}_{3k}' = \bar{r}_{3k}$, we may take $\Sigma r_{3k}' = r_{3b}(B_1 + B_3)$. The use of the above formulæ is therefore tantamount to supposing that the discrepancies between the true and the observed values for the several correlations (i.e., the residuals left after all the factors have been extracted) are the result of chance fluctuations, and will therefore sum to zero within the columns or rows (or sections of them) over which the summation extends, provided a sufficient number of tests has been used.

Since, in most psychological investigations, the number of tests on which the factors are based is decidedly small (particularly in the case of the group factors), this assumption is only very approximately true. Hence for any rigorous determination it would be better to proceed by seeking

---

[1] For the multiplication of partitioned matrices see Cullis, W. E. (1918), *Matrices and Determinoids*, II, pp. 6f., or Aitken, A. C. (1939), *Determinants and Matrices*, pp. 21–25.

TABLE XXII. CONSTRUCTION OF CORRELATION TABLE FOR ONE GROUP FACTOR

Test	1	2	Sum	3	4	Sum	Total
1	$r_{1b}^2 + r_{1c}^2$	$r_{1b}r_{2b} + r_{1c}r_{2c}$	$r_{1b}B_1 + r_{1c}C_1$	$r_{1b}r_{3b}$	$r_{1b}r_{4b}$	$r_{1b}B_3$	—
2	$r_{2b}r_{1b} + r_{2c}r_{1c}$	$r_{2b}^2 + r_{2c}^2$	$r_{2b}B_1 + r_{2c}C_1$	$r_{2b}r_{3b}$	$r_{2b}r_{4b}$	$r_{2b}B_3$	—
Sum	$r_{1b}B_1 + r_{1c}C_1$	$r_{2b}B_1 + r_{2c}C_1$	$B_1^2 + C_1^2$	$r_{3b}B_1$	$r_{4b}B_1$	B_1B_3	—
3	$r_{3b}r_{1b}$	$r_{3b}r_{2b}$	$r_{3b}B_1$	r_{3b}^2	$r_{3b}r_{4b}$	$r_{3b}B_3$	$r_{3b}(B_1 + B_3)$
4	$r_{4b}r_{1b}$	$r_{4b}r_{2b}$	$r_{4b}B_1$	$r_{4b}r_{3b}$	r_{4b}^2	$r_{4b}B_3$	$r_{4b}(B_1 + B_3)$
Sum	$r_{1b}B_3$	$r_{2b}B_3$	B_3B_1	$r_{3b}B_3$	$r_{4b}B_3$	B_3^2	—
Total	—	—	—	$r_{3b}(B_1 + B_3)$	$r_{4b}(B_1 + B_3)$	—	$B_3(B_1 + B_3)$

those particular values for the saturations which, with any postulated factor-pattern,[1] will maximize the likelihood of the test-measurements or test-correlations actually observed. This leads to a procedure which minimizes the square-sums of the residuals, instead of assuming that their algebraic sum will be zero. In practice, however, it is usually found that the differences in the results obtained by the two methods are not large enough to affect any demonstration of the existence or essential nature of the factors so discovered.[2]

II. *Calculation of Rotation Matrix*

Let B be the summed bipolar-matrix, G the summed group factor matrix, and T the required rotation matrix. Then $BT = G$, and $GT' = B$. Writing the last equation explicitly and substituting the known values for G and B from Table XVIII A and C, we have

$$\begin{bmatrix} \cdot 8816 & \cdot 0579 & \cdot 0000 & \cdot 0000 \\ 1\cdot 7389 & \cdot 6299 & \cdot 0000 & \cdot 0000 \\ 1\cdot 5875 & \cdot 0000 & \cdot 8244 & \cdot 0000 \\ 1\cdot 7134 & \cdot 0000 & \cdot 0000 & 1\cdot 4228 \end{bmatrix} \begin{bmatrix} a & e & j & p \\ b & f & k & q \\ c & g & m & r \\ d & h & n & s \end{bmatrix} = \begin{bmatrix} \cdot 8506 & \cdot 1568 & \cdot 1062 & \cdot 1453 \\ 1\cdot 7352 & \cdot 4177 & \cdot 4628 & -\cdot 1453 \\ 1\cdot 6300 & \cdot 4680 & -\cdot 5690 & \cdot 0000 \\ 1\cdot 9681 & -1\cdot 0425 & \cdot 0000 & \cdot 0000 \end{bmatrix}$$

where lower case letters are substituted for the several columns of T'. The above matrix equation provides us with four systems of simultaneous linear equations, each system containing four equations to be solved for four unknown quantities. Let us take the first column of unknown quantities, namely, a, b, c, and d. Changing to the ordinary notation, the matrix equation for these quantities may be expanded as follows :

$$\begin{array}{lllll} \cdot 8816a + \cdot 0579b & & & = \cdot 8506 & \quad . \quad . \quad . \quad \text{(i)} \\ 1\cdot 7389a + \cdot 6299b & & & = 1\cdot 7352 & \quad . \quad . \quad . \quad \text{(ii)} \\ 1\cdot 5875a & + \cdot 8244c & & = 1\cdot 6300 & \quad . \quad . \quad . \quad \text{(iii)} \\ 1\cdot 7134a & & + 1\cdot 4228d & = 1\cdot 9681 & \quad . \quad . \quad . \quad \text{(iv)} \end{array}$$

Multiplying (i) by $\cdot 6299/\cdot 0579 = 10\cdot 8791$, we have :

$$9\cdot 5910a + \cdot 6299b = 9\cdot 2538 \quad . \quad . \quad . \quad \text{(v)}$$

Subtracting (ii) from (v) we get :

$$7\cdot 8521a = 7\cdot 5186.$$

$$\text{Hence} \quad a = \cdot 9575.$$

Substituting for a, we obtain

$$\begin{array}{ll} \text{from (ii) } 1\cdot 6651 + \cdot 6299b = 1\cdot 7352. & \text{Hence } b = \cdot 1113. \\ \text{from (iii) } 1\cdot 5200 + \cdot 8244c = 1\cdot 6300. & \text{Hence } c = \cdot 1334. \\ \text{from (iv) } 1\cdot 6405 + 1\cdot 4228d = 1\cdot 9681. & \text{Hence } d = \cdot 2303. \end{array}$$

The figures thus obtained form the first row of T in Table XIX A. We can solve the equations for the three remaining columns of T' in the same way. It will be noted that the multiplier for the first equation (10·8791) is the same for all four columns. Hence in practice it is quicker to calculate corresponding steps for the four columns at the same time.

When all the values for the rotation matrix have been calculated by this method, the figures should be checked by verifying that the square-sum for each of the four rows amounts to 1·0000, and the product-sum for every pair of rows to zero.

I should like to express my indebtedness to my colleagues, Dr. Banks and Mr. Summerfield, who have been good enough to read this article in proof, and check many of the calculations.

---

[1] I here use the phrase 'factor-pattern' to denote the general scheme of positive, negative, or zero saturations, regardless of their numerical value. The full procedure would require us to calculate the amounts of the maximum likelihoods obtainable with *different* hypothetical factor-patterns (e.g., a two-factor pattern, a general-plus-bipolar factor-pattern, a basic-plus-group factor-pattern, and a 'simple structure' of group factors only), and then adopt, as provisionally proved, that hypothesis which furnished the largest value for its maximum likelihood—a proof by 'maximal maximum.'

[2] More detailed proofs for these various cases are given in the roneo'd *Laboratory Notes on Factor Analysis : Group Factor Method*. For a fuller discussion of the initial definitions and postulates see 7, pp. 246f., which incorporates the essentials from earlier memoranda.

FACTOR ANALYSIS BY MAXIMUM LIKELIHOOD: A CORRECTION

By D. N. LAWLEY

The Mathematical Institute, University of Edinburgh

In a recent paper (1) Mr. W. G. Emmett has quoted formulæ for the standard errors of factor loadings estimated by my method of maximum likelihood. The use of the formulæ, however, requires an additional qualification, since the derivation given in the paper (2) from which they were quoted is itself in need of correction.

It has now been found that the estimated loadings obtained for a given set of tests in one factor are slightly correlated with estimated loadings obtained for the same set of tests in another factor. This in turn means that the standard errors given by the formula for the estimated loadings in the rth factor are correct only if the values of loadings in previous factors are regarded as already given.

I should add that this correction in no way detracts from the value of Mr. Emmett's own paper.

REFERENCES

1. Emmett, W. G. (1949). 'Factor Analysis by Lawley's Method of Maximum Likelihood. *Brit. J. Psych. Stat. Sect.*, 90–97.
2. Lawley, D. N. (1949). 'Problems in Factor Analysis.' *Proc. Roy. Soc. Edin.* *A*, LXII, 1–6.

TESTS OF SIGNIFICANCE IN FACTOR ANALYSIS

By M. S. BARTLETT

Department of Mathematics, University of Manchester

I. *Description and Illustration of a Test of Significance for Analysis into Principal Components.* II. *Remarks on Lawley's Maximum Likelihood Method.* III. *Further Theoretical Discussion.*

I. DESCRIPTION AND ILLUSTRATION OF A TEST OF SIGNIFICANCE FOR ANALYSIS INTO PRINCIPAL COMPONENTS

Introductory Remarks. In a previous paper (1) in this *Journal* I distinguished between two types of statistical analysis, external canonical or factor analysis between two groups of variables and internal factor analysis of a single group. By the latter was meant the particular statistical analysis of the internal correlation structure known more familiarly as an ' analysis into principal components.' It was shown that the appropriate tests of significance of the ' factors ' emerging from the statistical analysis were quite distinct in the two cases ; but while a fairly satisfactory test was available for the external case, no analogous test was given in the internal case, only a method of assessing the significance of the smallest root but one against the smallest. It is the main purpose of the present paper to remedy this deficiency, and describe a more general test analogous to the test developed for the external case.

It should be noticed that an analysis into principal components is distinct from an analysis corresponding to some postulated psychological pattern, and has a different interpretation. The most systematic attempt to cover the latter problem is perhaps that by Lawley (2, 3, 4), using the method of maximum likelihood. A critical discussion of this method is given in Part II, where it is shown further that the associated χ^2 test proposed by Lawley in 1940 has a partial connexion with the χ^2 technique proposed in I. For convenience the theoretical justification of the formulæ and methods described in I (and, to some extent, in II) is deferred to Part III, which may be omitted by the non-mathematical reader.

The present paper does not entirely clear up the problems which it raises for either form of analysis ; but the procedures discussed seem sufficiently justified to be worth reviewing at this stage, even although further theoretical investigation is still needed.

The χ^2 Approximation. Let us denote by λ_i $(i = 1, 2, \ldots, p)$ the p roots, in descending order of magnitude, of the determinantal equation

$$| R - \lambda | = 0, \qquad (1)$$

where R is a correlation matrix of p variables. It should be noted that for the determinant $| R |$ we have the relation $| R | = \lambda_1 \lambda_2 \ldots \lambda_p$. Let the total number of degrees of freedom available from the original observations be n ; we shall usually have $n = r - 1$, when deviations from the general mean are employed for each variable, where r is the total number of observations for each variable, i.e., in the present context r is the total number of persons tested.[1] The statistical significance of the entire correlation structure may be tested if necessary by calculating the quantity[1]

---

[1] Note that the number of degrees of freedom s of Table I of the previous paper was defined as $r - 2$. This is the usual convention for the test of significance of a single correlation, but the degrees of freedom $n = r - 1$ for each variable is more appropriate here.

$$\chi^2 = -\{n - \tfrac{1}{6}(2p + 5)\} \log_e | R | \ , \qquad (2)$$

with $\frac{1}{2}p(p - 1)$ degrees of freedom. However, the usual situation will be that after the removal of the largest roots we require to test the significance of those left. The χ^2 test may still, as in an external analysis, be used for this purpose, provided that the roots removed correspond to well-determined (i.e., highly significant) factors. We now take

$$\chi^2 = -\{n - \tfrac{1}{6}(2p + 5) - \tfrac{2}{3}k\} \log_e R_{p-k} \, , \qquad (3)$$

with $\frac{1}{2}(p - k)(p - k - 1)$ degrees of freedom, after k roots $\lambda_1 \ \lambda_2 \ldots \lambda_k$ have been determined, where

$$R_{p-k} = | R | \Bigg/ \left\{ \lambda_1 \lambda_2 \ldots \lambda_k \left[\frac{p - \lambda_1 - \lambda_2 \ldots - \lambda_k}{p - k} \right]^{p-k} \right\} . \qquad (4)$$

It will be noticed that the test is formulated to indicate the significance of the residual roots ; however, it is often convenient to present a complete χ^2 table, provided this is properly interpreted. Since the multiplying factor in (3) changes with k, the complete table is not strictly additive unless a constant factor is substituted ; we might employ the smallest, corresponding to $k = p - 2$: the factor is then $n - p + \frac{1}{2}$. But if n is not large, the full advantage of the approximation is retained if the proper factor is used for each individual test.

Numerical Illustrations. Two examples will be given to illustrate this test. For the first I have taken the data of Truman Kelly's quoted in Hotelling's original discussion (5) of an analysis into principal components, consisting of four tests (reading speed, reading power, arithmetic speed, and arithmetic power) carried out on 140 children. For the matrix R we have[2]

$$\begin{pmatrix} 1 & 0\cdot698 & 0\cdot264 & 0\cdot081 \\ 0\cdot698 & 1 & -0\cdot061 & 0\cdot092 \\ 0\cdot264 & -0\cdot061 & 1 & 0\cdot594 \\ 0\cdot081 & 0\cdot092 & 0\cdot594 & 1 \end{pmatrix}$$

The variances taken up by the four roots were

$$\begin{array}{rcl} \lambda_1 & = & 1\cdot846 \\ \lambda_2 & = & 1\cdot465 \\ \lambda_3 & = & 0\cdot521 \\ \lambda_4 & = & 0\cdot167 \\ \hline \text{Total} & = & 3\cdot999 \end{array} \qquad \begin{array}{rcl} | R | & = & \lambda_1 \lambda_2 \lambda_3 \lambda_4 \\ & = & 0\cdot2353 \end{array}$$

For the quantities R_{p-k} in (3) we find

$$\begin{array}{rcl} R_2 & = & 0\cdot521 \times 0\cdot167 \times (2/0\cdot689)^2 = 0\cdot7330 \\ R_3 & = & (0\cdot2353/1\cdot846) \times (3/2\cdot154)^3 = 0\cdot3444 \\ R_4 & = & | R | = 0\cdot2353 \ . \end{array}$$

In practice when the smallest roots have not been evaluated $| R |$ must of course be evaluated directly. Using the common factor $n - p + \frac{1}{2} = 139 - 4 + \frac{1}{2} = 135\cdot5$, we obtain the additive χ^2 analysis :

---

[1] As noted in Part III, the criterion $| R |$ was considered by Wilks in 1932 ; the χ^2 approximation proposed in equation (2) is closely connected with the χ^2 approximation I have put forward for external analysis (see also Box (11) p. 344).

[2] In order to make use of Hotelling's analysis for this first example these correlations, corrected for attenuation, have been used, though of course this correction will affect test of significance to some (unknown) extent which is here ignored.

	D.F.	χ^2
λ_1	3	$-135{\cdot}5\,(\log_e 0{\cdot}2353 - \log_e 0{\cdot}3444) = 51{\cdot}6$
λ_2	2	$-135{\cdot}5\,(\log_e 0{\cdot}3444 - \log_e 0{\cdot}7330) = 102{\cdot}3$
λ_3	1	$-135{\cdot}5\,(\log_e 0{\cdot}7330) = 42{\cdot}1$
λ_4	—	
Total ..	6	$-135{\cdot}5\,(\log_e 0{\cdot}2353) = 196{\cdot}0$

The best approximation for the total χ^2 would actually be $-\{n - \frac{1}{6}(2p + 5)\} \log_e |R| = -136{\cdot}833\,(\log_e 0{\cdot}2353) = 198{\cdot}0$, but of course all the components are so highly significant that this difference is unimportant. It may be wondered why the second largest component gives an apparently more significant χ^2 than the first; but it should be remembered that logically the significance of at least one root is to be judged from the total χ^2, not the component corresponding to λ_1. The apparent significance of this component according to the above table represents the magnitude of the root compared with the variation still remaining, and the latter contains quite a large second component. On the other hand, the apparent significance of the latter according to the analysis of the table is judged in relation to the residual variation after this component has in turn been removed, and this residual variation is comparatively small.

As the second example I have taken the hypothetical data discussed by Burt (6), for which the number of persons is only 6. It is indicated in the theoretical discussion in III that we may expect the test presented here to work satisfactorily, at least for the total χ^2, for as low as about 10 degrees of freedom.[1] This means that the test must be used with considerable caution in the present example, which is unrealistic in its very small number of persons. In view of the small value of n, the χ^2 table for this example will be presented not in additive form, but in the form in which it would arise as each set of residual factors is tested in turn for significance, the best multiplying factor given in (3) being used in each case.

From Table XIII of Burt's paper we have

$$\begin{aligned} \lambda_1 &= 3{\cdot}1021 \\ \lambda_2 &= 0{\cdot}7328 \\ \lambda_3 &= 0{\cdot}1593 \\ \lambda_4 &= 0{\cdot}0058 \\ \hline \text{Total} &= 4{\cdot}0000 \end{aligned} \qquad \begin{aligned} |R| &= \lambda_1 \lambda_2 \lambda_3 \lambda_4 \\ &= 0{\cdot}002101 \end{aligned}$$

Hence

$$\begin{aligned} R_2 &= 0{\cdot}1593 \times 0{\cdot}0058 \times (2/0{\cdot}1651)^2 = 0{\cdot}1356 \\ R_3 &= (0{\cdot}002101/3{\cdot}1021) \times (3/0{\cdot}8979)^3 = 0{\cdot}02526 \\ R_4 &= |R| = 0{\cdot}002101 . \end{aligned}$$

The χ^2 test for the residual roots is thus:

	D.F.	χ^2	$P = 0{\cdot}05$ level
$\lambda_1\ \lambda_2\ \lambda_3$	3+2+1	$-2{\cdot}833\,(\log_e 0{\cdot}002101) = 17{\cdot}47$	12·59
$\lambda_2\ \lambda_3$	2+1	$-2{\cdot}167\,(\log_e 0{\cdot}02526) = 7{\cdot}97$	7·82
λ_3	1	$-1{\cdot}5\,(\log_e 0{\cdot}1356) = 3{\cdot}00$	3·84

---

[1] By these degrees of freedom is meant of course the number depending on the number of persons, *not* the degrees of freedom for χ^2, depending only on the number of tests.

Even with such a small sample we are thus able to conclude that the first two factors are significant ; the third, however, does not reach significance (cf. the conclusion in the final paragraph of my previous paper (1)).

Further Comments. It need hardly be stressed that I am concerned here with *statistical* significance. It does not follow that all the factors which reach statistical significance in a large sample necessarily remove a very large fraction of the variance ; and hence some of them may be comparatively unimportant in practice. Again, even if they are numerically important, this has no necessary implications of psychological or other reality of the factors. Merely the correlation structure of the variables is being investigated in its relation to variance : for this reason no significance can ever be attached to the last root, for it would be equivalent to asking for the correlation structure of a single variable.

If of course the basic equation (1) is replaced by another, then the above tests become inappropriate ; for example, if we ask for the significance of the factors of an equation like

$$| R - \lambda e_{ii} | = 0 , \tag{5}$$

where e_{ii} denotes error variances (assumed known), then, as I showed in (1), the analysis is a special case of an external factor analysis, and other tests become relevant. Or if in place of the empirical analysis of equation (1) we substitute a factor analysis of more orthodox psychological type, then, as noted in the introductory remarks, other tests again will be required. Some discussion is given in Part II of a factor analysis of standard type, which in vector and matrix notation is represented by the equation

$$T = M_0 F_0 + M_1 F_1 , \tag{6}$$

where T represents the vector of standardized tests, F_0 the set of (standardized) general or group factors, F_1 the (standardized) specifics, and (M_0 , M_1) the factor 'loadings.' Since the difference between the factors of equation (6) and those corresponding to the analysis (1) is not always realized, it may be advisable to illustrate this difference by a simple example. Consider three tests of which the first two are correlated with coefficient ρ (positive), and the third is entirely independent. On the basis of equation (6) this might be interpreted as indicating one 'general' factor common to the first two tests, but with zero loading for the third, the remaining factors being specific to each test. On the basis of an analysis into principal components, however, it is easily seen that the roots are $1 + \rho$, 1 and $1 - \rho$, and it will be noticed that in order of contribution to total variance it is the second root which corresponds to the uncorrelated test and, for any reasonable size of sample, this will be significant compared with the last. Thus once significant correlation is present it may automatically cause all the roots to be distinct in a principal components analysis, which should not be assumed to lead necessarily to the *simplest* factor-pattern.

II. REMARKS ON LAWLEY'S MAXIMUM LIKELIHOOD METHOD

The Maximum Likelihood Method in Factor Analysis. To determine the structural factor equation of the type represented by equation (6), Lawley has investigated two methods based on the principle of maximum likelihood. The first, which he has called 'Method I,' is based on the variance-covariance sampling distribution of the test scores ; the second, 'Method II,' is based more directly on the individual test scores (see his paper (3)). At first sight the second method would appear the more fundamental, but it is known to lead to difficulties. In a recent paper (7) M. G. Kendall raised doubts on the suitability of the maximum likelihood method in the present context ; and Dr. Lawley has informed me that he had already entertained similar doubts, at least as regards 'Method II.' In the discussion following Kendall's paper, I pointed out that the factor equation (6) is equivalent to another set of structural equations (familiar in econometrics) for which it is definitely known that the standard maximum likelihood method breaks down.

The question remains of the validity and status of Lawley's 'Method I,' and a detailed re-examination and formulation of the most appropriate statistical procedure for dealing with the structural equation (6) seems still required. Some provisional comments on Lawley's first method may, however, be useful.

Lawley's 'Method I,' and its Associated χ^2 Test. I was at first sceptical of the value of this method because of its obvious breakdown in elementary cases. For example, it does not work for the case of two tests and one general factor, for we should then have three observed quantities, the two variances and the covariance, and four unknown loading coefficients (or equivalently, when the tests have been standardized, one observed correlation, and two unknown loading coefficients for the general factor) ; in other fields this case often arises in the guise of fitting a straight line to the plot of two variables both of which are subject to error (see, for example, Bartlett (8)).

In psychology the number of tests is usually large enough for this difficulty not to arise. The validity of the maximum likelihood method applied to the variance-covariance distribution then depends (i) on our deciding to consider only the information contained in the variance-covariance (or equivalent correlation) matrix, and (ii) on Lawley's procedure, given (i), providing efficient estimates. Since the general validity of the maximum likelihood approach can no longer be assumed, (ii) is no longer automatic, and appears to require demonstration. One possible argument is to note that enough factors and unknown coefficients can always be introduced to take up all the available degrees of freedom and yet avoid redundancy ; representation of the correlation matrix by still fewer parameters merely implies further linear restrictions among the unknown coefficients, and this should hardly affect the question of efficiency when the likelihood method is used.

But it might be noticed that the method depends on the correlation matrix being compatible with the postulated factor-pattern. For example, with only three tests it is not possible to have a non-redundant factor equation (6) except in the case of only one general factor, which should take up all the degrees of freedom. Yet the correlation matrix

$$\begin{pmatrix} 1 & 0{\cdot}60 & -0{\cdot}28 \\ 0{\cdot}60 & 1 & 0{\cdot}60 \\ -0{\cdot}28 & 0{\cdot}60 & 1 \end{pmatrix}$$

is only compatible with a minimum of two general factors. This would have to be revealed by equations impossible to satisfy with real coefficients, and not by Lawley's associated χ^2 test. If the above matrix arose in a sample, it is not clear how we should test the goodness of fit of one general factor. If we ignore such difficulties, Lawley's method appears to provide a satisfactorily objective procedure for dealing with equation (6).

It has been stressed that this factor analysis is quite distinct from that discussed in Part I of this paper ; but it is an attractive link between the two analyses that the total χ^2 corresponding to the significance of the unreduced correlation matrix is necessarily the same, and only because of the difference between the factors extracted in the two analyses does the analysis of the total χ^2 into its respective components differ.

It follows (see the last section of Part III of this paper) that the form of Lawley's test which should be more exact for moderate-sized samples is the determinantal form

$$\chi^2 \sim -n' \log_e |R| / |\hat{R}| ,$$

where $|\hat{R}|$ denotes the estimated correlation matrix when k general factors (and p specifics) have been fitted, and the multiplying coefficient n' be taken, not to be n, but the coefficient $n - \frac{1}{6}(2p + 5)$ used in (2), or perhaps even better, the coefficient $n - \frac{1}{6}(2p + 5) - \frac{2}{3}k$ used in (3). The number of degrees of freedom of χ^2 is $\frac{1}{2}(p - k)(p - k - 1) - k$ (subject, of course, to the difficulties raised above).

III. FURTHER THEORETICAL DISCUSSION

Equivalent Analyses of Correlation Structure. Any resemblance of the test described in I above to the test for external analysis described in the earlier paper is more than superficial. Let us consider the logical relations among a group of variables, say, four for definiteness. The possible correlations may be considered in more than one way. Thus we may consider the correlation between variables x_1 and x_2, then the relation between x_3 and the group x_1, x_2, then between x_4 and the group x_1, x_2, x_3. As another alternative, we might consider the external relation between the two groups x_1, x_2 and x_3, x_4; then the further internal relations between x_1 and x_2 in the first group, and between x_3 and x_4 in the second. The internal relations in a single group can thus be regarded as a number of 'external' relations between various subgroups. Hence a test of the internal relations can be built up from the tests of the equivalent external relations between sub-groups. If we carry out this theoretical programme according, say, to the first of the alternative analyses mentioned above, where the relation of each further variable is considered successively, we arrive from the theory of the χ^2 test for external analysis (9) at the successive χ^2 quantities:

$$\begin{array}{ll}
-\{n - \frac{1}{2}(3)\} \log_e (1 - r_{12}^2) , & (1 \text{ d.f.}) , \\
-\{n - \frac{1}{2}(4)\} \log_e (1 - R_{3\,,\,12}^2) , & (2 \text{ d.f.}) , \\
\vdots & \\
-\{n - \frac{1}{2}(p + 1)\} \log_e (1 - R_{p\,,\,12 \cdots p-1}^2) , & (p - 1 \text{ d.f.}) ,
\end{array}$$

where r_{12} is the observed correlation between x_1 and x_2, $R_{p\,,\,12 \cdots p-1}$ the multiple correlation of x_p with $x_1, x_2, \ldots, x_{p-1}$. In order to make any test arrived at invariant to the route of arrival, it is obviously convenient to choose a mean multiplying factor, which (weighting with the degrees of freedom) becomes $n - \frac{1}{2}f$, where

$$\tfrac{1}{2}p(p - 1)f = (1 \times 3) + (2 \times 4) + (3 \times 5) + \ldots (p - 1)(p + 1) ,$$

or

$$f = \tfrac{1}{3}(2p + 5) .$$

It is well known that

$$(1 - r_{12}^2)(1 - R_{3\,,\,12}^2) \ldots (1 - R_{p\,,\,12 \cdots p-1}^2) = |R| ,$$

whence we obtain as an approximate χ^2, with $1 + 2 + \ldots + (p - 1) = \frac{1}{2}p(p - 1)$ degrees of freedom, the expression proposed in equation (2).

Direct Derivation of the χ^2 Approximation. The above method of approach was the one I used first to arrive at (2); but it is satisfactory that the test finally arrived at depends on the correlation determinant $|R|$, a criterion considered many years ago by Wilks (10). It is useful to check directly the appropriateness of the χ^2 approximation by writing down the moment-generating function of $-n \log_e |R|$, by suitably adapting Wilks' formula for the sth order moment of $|R|$ (cf. the derivation in (9) of the χ^2 test for an external analysis). Alternatively the moment-generating function could be written down from the simultaneous distribution of the correlations R, or from the independence of the components listed above. We obtain

$$M(t) = \frac{\Gamma^{p-1}(\tfrac{1}{2}n) \prod_{i=1}^{p-1} \Gamma\{\tfrac{1}{2}(n-i) - nt\}}{\Gamma^{p-1}(\tfrac{1}{2}n - nt) \prod_{i=1}^{p-1} \Gamma\{\tfrac{1}{2}(n-i)\}}, \tag{7}$$

and it is a matter of straightforward algebra coupled with Stirling's approximation formula for a Γ-function to show that

$$\begin{aligned} M(t) &= \sum_{i=1}^{p-1}\left\{ t\left[i + \frac{i^2+2i}{2n} + \dots\right] + t^2\left[i + \frac{i^2+2i}{n} + \dots\right] + \frac{4}{3}t^3\left[i + \frac{3(i^2+2i)}{2n} + \dots\right] + \dots\right\} \\ &= \tfrac{1}{2}p(p-1)\,t\left[1 - \frac{1}{6n}(2p+5)\right] + \tfrac{1}{2}p(p-1)\,t^2\left[1 - \frac{1}{6n}(2p+5)\right]^2 \\ &\quad + \tfrac{1}{2}p(p-1)\frac{8t^3}{3!}\left[1 - \frac{1}{6n}(2p+5)\right]^3 + \dots \end{aligned} \tag{8}$$

to $0\left(t^3, \frac{1}{n}\right)$. Hence the improved approximation is

$$\chi^2 \sim -\{n - \tfrac{1}{6}(2p+5)\}\log_e |R|,$$

confirming the mean multiplying factor arrived at by the previous argument, and from (8) having $\tfrac{1}{2}p(p-1)$ degrees of freedom.[1]

Closeness of the χ^2 Approximation. The order of closeness of this approximation will clearly be comparable to that for an external analysis, in view of their relationship, and will consequently be expected to work quite well down to a sample with about 10 degrees of freedom (9). Let us, for example, in the case of only two variables, where merely the test of a single correlation is involved, check the closeness of the approximation. We find for χ^2, corresponding to the correct critical $P = 0{\cdot}05$ or $0{\cdot}01$ value of r_{12}, the values :

	$P = 0{\cdot}05$	$P = 0{\cdot}01$
$n = 10$	3·84	6·60
$n = 20$	3·84	6·63
$n = \infty$	3·841	6·635

Effect of Eliminating the Larger Roots. The approximate effect of eliminating the larger roots in the factor analysis has next to be considered. To some extent it is possible to give an argument analogous to that used for an external analysis, when the roots eliminated are well determined. In this case the orthogonal directions corresponding to these roots in the geometrical representation of the variables in the n-dimensional sample space will be well determined, and we may consider the analysis of the remaining $p - k$ roots in a space of $n - k$ dimensions (if k roots have been eliminated).

The adjustment of the multiplying factor gives

$$(n-k) - \tfrac{1}{6}\{2(p-k)+5\} = n - \tfrac{1}{6}(2p+5) - \tfrac{2}{3}k,$$

as given in (3). The expression for R_{p-k} given in (3) is merely a convenient method of obtaining the expression in the remaining roots analogous to the determinant $|R| = \lambda_1 \lambda_2 \dots \lambda_p$, when we are given $|R|$ and $\lambda_1 \lambda_2 \dots \lambda_k$. It should be remembered that significance in an internal analysis can only be on the basis of the relative values of the roots, which must after each elimination be in effect re-scaled to unit mean variance. This explains the factor $[(p-k)/(\lambda_{k+1} + \dots \lambda_p)]^{p-k}$ in R_{p-k}, which may thus be written

$$\frac{|R|}{\lambda_1 \lambda_2 \dots \lambda_k}\left\{\frac{p-k}{p - \lambda_1 - \lambda_2 \dots - \lambda_k}\right\}^{p-k}.$$

[1] An independent and prior mention of this approximation is included in the recent comprehensive paper on such approximations by Box (11).

It must be admitted that the above justification of the proposed test for the residual roots is not as complete as could be wished. The problem would be slightly simpler if we were alternatively considering the analysis of test scores known to have true equal variances, but standardized to unity only for the *mean* variance. In this case the above form of the test for the residual roots can be given further justification, but of course the total χ^2 now has $p - 1$ more degrees of freedom, these representing the test of homogeneity of the p test variances about their unit mean variance. Correspondingly, the total χ^2 for the last $p - k$ factors would have (approximately) $\frac{1}{2}(p - k - 1)(p - k + 2)$ degrees of freedom instead of $\frac{1}{2}(p - k)(p - k - 1)$. One point that remains in some doubt pending a more detailed examination is whether the reduction in degrees of freedom that ensues from the individual standardization of the tests is automatically felt in the residual factor components (an assumption implicit in the proposed test), or is mainly absorbed by the larger roots.

The Alternative Forms of Lawley's χ^2 Test. In his 1940 paper (2), Lawley gave two alternative forms for the large-sample χ^2 test he developed for use with his maximum likelihood method (Method I). One was an arithmetically convenient sum of squares, but it is worth pointing out that the equivalent determinant formula arises more exactly than he perhaps emphasized, and should therefore be the more accurate one for moderate-sized samples.

To demonstrate this, it is convenient to define a 'homogeneous' likelihood function from the variance-covariance sampling distribution by its logarithm L, where

$$-2L + \text{constant} = n \log |C| - n \log |A| + n \operatorname{trace}(C^{-1}A) . \quad (9)$$

In this expression, where C denotes the true variance-covariance matrix and A the corresponding sample matrix, $-2L$ tends to χ^2 with $\frac{1}{2}p(p + 1)$ degrees of freedom as n increases. To determine the constant in (9) we substitute the estimates A for C, giving

$$\text{constant} = n \operatorname{trace}(A^{-1}A) = np . \quad (10)$$

If we alternatively substitute estimates of C from (efficient) factor estimates, with q independent unknown coefficients, we obtain

$$-2\hat{L} + \text{constant} = n \log |\hat{C}| - n \log |A| + n \operatorname{trace}(\hat{C}^{-1}A) , \quad (11)$$

with $\frac{1}{2}p(p + 1) - q$ degrees of freedom.

From (10) and (11),

$$\begin{aligned} -2\hat{L} &= -n \log |A| / |\hat{C}| + n \operatorname{trace}(\hat{C}^{-1}A) - np \\ &= -n \log |A| / |\hat{C}| + n \operatorname{trace}(\hat{C}^{-1}A - \hat{C}^{-1}\hat{C}) \\ &= -n \log |A| / |\hat{C}| , \end{aligned} \quad (12)$$

the last term vanishing when $\hat{C}$ corresponds to the maximum likelihood estimates (see Lawley (2)).

If no general factors are fitted, the estimates $\hat{C}$ consist simply of the specific factor variances coinciding with the test variances, and (12) reduces to $-n \log |R|$. It is then, as we should expect, identical with the total χ^2 proposed in the principal components analysis, and the modified multiplying coefficient $n - \frac{1}{6}(2p + 5)$ is consequently more accurate. When k general factors have also been fitted, (12) may be expressed in terms of the ratio of the observed and estimated *correlation* determinants. The more accurate multiplying coefficient $n - \frac{1}{6}(2p + 5)$ should still be more appropriate than the crude coefficient n, and the further modified coefficient $n - \frac{1}{6}(2p + 5) - \frac{2}{3}k$ would have some justification here also, though this last refinement is of course not very firmly established. The number of degrees of freedom comes out at $\frac{1}{2}(p - k)(p - k - 1) - k$, the number lost per component being one more than in the principal components analysis.

Finally, it might be noticed that an alternative analysis of the total χ^2 along the lines given in the first section of Part III would sometimes be of practical interest, and since, as was there pointed out, it would then rest on the known χ^2 approximation for *external* analyses, such an analysis would have a somewhat more precise justification than it has yet been found possible to give either for the principal components analysis or for the maximum likelihood factor analysis.

REFERENCES

1. Bartlett, M. S. (1948). 'Internal and external factor analysis.' *Brit. J. Psych., Stat. Sect.*, I, 73.
2. Lawley, D. N. (1940). 'The estimation of factor loadings by the method of maximum likelihood.' *Proc. Roy. Soc. Edin.*, IX, 64.
3. Lawley, D. N. (1942). 'Further investigations in factor estimation.' *Proc. Roy. Soc. Edin.*, LXI, 176.
4. Lawley, D. N. (1949). 'Problems in factor analysis.' *Proc. Roy. Soc. Edin.*, LXII, 1.
5. Hotelling, H. (1933). 'Analysis of a complex of statistical variables into principal components.' *J. Educ. Psych.*, XXIV, 417 and 498.
6. Burt, C. (1947). 'Factor analysis and analysis of variance.' *Brit. J. Psych., Stat. Sect.*, I, 3.
7. Kendall, M. G. (1950). 'Factor analysis as a statistical technique.' Symposium on factor analysis to be published in *J. Roy. Statist. Soc.* (Series B).
8. Bartlett, M. S. (1949). 'Fitting a straight line when both variables are subject to error.' *Biometrics*, 5, 207.
9. Bartlett, M. S. (1938). 'Further aspects of the theory of multiple regression.' *Proc. Camb. Phil. Soc.*, XXXIV, 33.
10. Wilks, S. S. (1932). 'Certain generalizations in analysis of variance.' *Biometrika*, XXIV, 471.
11. Box, G. E. P. (1949). 'A general distribution theory for a class of likelihood criteria.' *Biometrika*, XXXVI, 317.

A METHOD OF STANDARDIZING GROUP-TESTS

By D. N. LAWLEY
University of Edinburgh

Introduction.—A method has been given by Thomson (1) for standardizing group-tests, by means of which the scores obtained by a complete year-group of children in a test can be converted into 'intelligence quotients' or I.Q.s. These I.Q.s are not true quotients; but are similar to Binet I.Q.s in that they make allowance for difference of age, and are distributed normally about a mean of 100. The standard deviation of group-test I.Q.s has for convenience been fixed at 15, to agree very roughly with that of Binet I.Q.s.

The method of standardization depends essentially on finding for each month of age the 5, 16, 50, 84, and 95 percentiles of the distribution of 'raw scores.' Straight lines are then fitted to obtain for each month the estimated value of each of the percentiles. This means that for any given age within the range considered we can estimate the raw scores corresponding to I.Q.s of (approximately) 75, 85, 100, 115, and 125. Scores corresponding to other I.Q.s may then be estimated by interpolation or extrapolation.

The conversion tables so constructed suffer from the disadvantage that raw scores, and not I.Q.s, are given in the body of the table. In general the I.Q. corresponding to a given age and score can only be found by inverse interpolation. It was for this reason that the author developed a somewhat different method of standardization. This has for some years been used to standardize Moray House group-tests, and has been found to give results very close to the former method, with considerably less calculation.

Description of Method.—In developing the method the main idea has been to make the calculations as simple as possible, without too great loss of accuracy. The efficiency, in a statistical sense, could doubtless be increased, but not without adding materially to the time required. Errors of estimation arising from the construction of the conversion table are in most cases negligible compared with the errors of measurement of any mental test, however good.

We shall suppose that data are available for a complete year-group of children in some area, and that the total number of cases is large, say, at least a thousand. It will be assumed that the age a of a child is measured in units of completed months. We shall thus have twelve score distributions, one for each value of a. We may further assume that there are approximately the same number in each month-group.

Let us denote by x the 'raw score' of any child; and let us suppose that the age a is measured from the mean of the year-group. It will then be presumed that the raw score x depends only on the age a and on the I.Q. y of the child, which is, however, unknown. Thus we may write $x = \varphi(y, a)$, a relation which may alternatively be expressed in the form $y = F(x, a)$.

Now it is clear that there is no logical necessity for $F(x, a)$ to be a simple function of x and a. In practice, however, we may usually assume, to a sufficiently good approximation, that $y = f(x) - \beta a$, where $f(x)$ is some function of x only, while β is a positive constant not dependent on either x or a. Furthermore, the I.Q. y is to be chosen in such a way that it will be distributed normally with a mean of 100 and a standard deviation of 15. Since

$$f(x) = y + \beta a \qquad (1)$$

it follows that, for a given value of a, $f(x)$ must be distributed normally with a mean of $(100 + \beta a)$ and a standard deviation of 15.

When a is not fixed, the distribution of $f(x)$ will not be normal, since that of a is not normal but rectangular. We may nevertheless treat $f(x)$ as being very nearly normally distributed, since the term βa is small; and as y and a are distributed independently, the total variance σ_f^2 of $f(x)$ is given by $\sigma_f^2 = \sigma_y^2 + \beta^2\sigma_a^2$, where σ_y^2 and σ_a^2 are the variances of y and a respectively. Putting $\sigma_y^2 = 225$ and $\sigma_a^2 = 12$ (since the range of a is 12), we thus have

$$\sigma_f^2 = 225 + 12\beta^2 . \tag{2}$$

It is first of all necessary to estimate the constant β. To do this we take a given value of x, say, x_0, and calculate for each value of a, i.e., for each month-group, the proportion p of children with scores less than x_0. For greatest accuracy it is desirable to choose the value of x_0 to be near the median of the total score distribution. We then find for each value of p a corresponding value of z, chosen in such a way that p is the probability of a normal variate with zero mean and unit variance lying within the range $-\infty$ to z. For this purpose Karl Pearson's (2) Table I will be found useful.

For each value of a the quantity $100+15z$ will now provide an estimate of the I.Q. y corresponding to a score of x_0. In order to obtain the relationship between z and a for constant x we fit a straight line to the twelve values of z by the method of least squares (with equal weights). The slope of this line when multiplied by -15 gives us an estimate of β.

We now consider the total distribution of raw scores found by pooling those for the twelve month-groups. The scores will usually be grouped in intervals of 5 or 10 marks, and for the score corresponding to the top end of each interval we find the proportion P of children in the whole year-group obtaining less than this. We may then, if we choose, plot P against x and draw a free-hand curve through the resulting points. In actual practice it is easier to employ some simple, non-graphical method of smoothing the values of P.

For a set of equally spaced values of x we thus obtain corresponding proportions P, which may then be converted into normal deviates Z in the same manner in which p was converted into z. For these values of x we now put

$$f(x) = 100 + \sigma_f Z , \tag{3}$$

where σ_f^2 is as given by equation (2). The mean and standard deviation of $f(x)$ will then be as previously specified. We may regard $f(x)$ as the I.Q. corresponding to a score x and an age given by $a = 0$, since in dealing with the total distribution of score we have in effect averaged out age. For values of x other than the end points of the grouping intervals $f(x)$ must be obtained by interpolation or extrapolation. In theory it might be possible to make some assumption regarding the mathematical form of $f(x)$. This approach to the problem has, however, not been followed owing to the difficulties which would arise in estimating the unknown parameters thereby involved.

To find the I.Q. y corresponding to any pair of values of x and a we use both equations (1) and (3). Hence

$$\begin{aligned} y &= f(x) - \beta a \\ &= (100 - \beta a) + \sigma_f Z . \end{aligned} \tag{4}$$

It will be found convenient to calculate the I.Q.s y_1 corresponding to the lowest value of a required, say, a_1. Thus

$$y_1 = (100 - \beta a_1) + \sigma_f Z . \tag{5}$$

For other values of a the I.Q. y is then given by

$$y = y_1 - \beta(a - a_1) . \tag{6}$$

Numerical Example. To illustrate the method of standardization a numerical example is given below. The data represent the score distributions of a complete year-group for a Moray House group-test. The ages are given in years and completed months at date of test, and range from 10 : 8 (10 years 8 months) to 11 : 7.

Two values of x_0 were chosen, namely, 49·5 and 79·5; these correspond to the sets of values p, z and p', z' respectively. The slopes of the two lines found by the method of least squares differ remarkably little; they are $-{\cdot}03524$ and $-{\cdot}03514$ respectively. We therefore estimate β as $15 \times {\cdot}0352 = 0{\cdot}528$. The value of σ_f is then given by $\sigma_f = \sqrt{225 + 12(0{\cdot}528)^2} = 15{\cdot}11$.

In constructing the conversion table it is inconvenient to use the Z values corresponding to the end points of the score intervals, namely, 9·5, 19·5, 29·5, etc. We therefore interpolate or extrapolate to obtain the values for $x = 10, 20, \ldots 90$. Using equation (5) we then find the I.Q.s y_1 for age 10 : 8. The difference between the lowest age 10 : 8 and the median age 11 : 1½ is 5·5 months.

TABLE I. DISTRIBUTION OF SCORE AND AGE

Score interval	Age in years and completed months at date of test												Totals	Cumul. Totals	P	Z
	10 : 8	10 : 9	10 : 10	10 : 11	11 : 0	11 : 1	11 : 2	11 : 3	11 : 4	11 : 5	11 : 6	11 : 7				
90–100	19	15	22	26	21	25	25	25	21	28	31	42	300	5368		
80– 89	31	40	36	30	45	37	47	34	39	51	46	60	496	5068	·944	1·589
70– 79	38 424	45 366	48 419	44 421	40 390	47 371	41 411	50 392	40 348	43 331	59 362	53 337	548	4572	·852	1·045
60– 69	38	36	55	52	48	67	53	52	41	49	54	48	593	4024	·750	·674
50– 59	46	37	46	57	51	53	58	47	45	44	50	39	573	3431	·639	·356
40– 49	57 302	39 248	41 270	60 268	41 251	44 204	59 259	48 243	47 222	47 195	38 199	34 197	555	2858	·532	·080
30– 39	53	41	63	45	65	29	47	59	45	39	40	44	570	2303	·429	−·179
20– 29	54	47	48	49	47	42	55	46	48	34	40	44	554	1733	·323	−·459
10– 19	64	60	56	57	54	45	49	46	46	40	49	40	606	1179	·220	−·772
0– 9	74	61	62	57	44	44	49	44	36	35	32	35	573	573	·107	−1·243
Totals . .	474	421	477	477	456	433	483	451	408	410	439	439	5368			
p	·637	·589	·566	·562	·550	·471	·536	·539	·544	·476	·453	·449				
z	·350	·225	·166	·156	·126	−·073	·090	·098	·110	−·060	−·118	−·128				
p'	·895	·869	·878	·883	·855	·857	·851	·869	·853	·807	·825	·768				
z'	1·254	1·122	1·165	1·190	1·058	1·067	1·041	1·122	1·049	·867	·935	·732				

TABLE II. CONVERSION TABLE

Age at Test Date	RAW SCORES																		
	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95
10 : 8	80	85	88	92	94	96	98	100	102	104	106	109	111	113	116	119	123	127	133
10 : 9	79	84	88	91	94	96	98	100	102	104	106	108	110	113	115	119	122	127	132
10 : 10	79	83	87	90	93	95	97	99	101	103	105	107	110	112	115	118	122	126	132
10 : 11	78	83	87	90	92	95	97	99	101	103	105	107	109	112	114	117	121	126	131
11 : 0	78	82	86	89	92	94	96	98	100	102	104	106	109	111	114	117	121	125	131
11 : 1	77	82	86	89	91	94	96	98	100	102	104	106	108	111	113	116	120	125	130
11 : 2	77	81	85	88	91	93	95	97	99	101	103	105	108	110	113	116	120	124	130
11 : 3	76	81	85	88	90	92	95	97	99	101	103	105	107	110	112	115	119	124	129
11 : 4	76	80	84	87	90	92	94	96	98	100	102	104	107	109	112	115	119	123	129
11 : 5	75	80	84	87	89	91	94	96	98	100	102	104	106	109	111	114	118	123	128
11 : 6	75	79	83	86	89	91	93	95	97	99	101	103	106	108	111	114	118	122	128
11 : 7	74	79	83	86	88	90	93	95	97	99	101	103	105	108	110	113	117	122	127

Hence, putting $a_1 = -5{\cdot}5$, we find that

$$y_1 = 100 - 0{\cdot}528 \times (-5{\cdot}5) + 15{\cdot}11\,Z$$
$$= 102{\cdot}90 + 15{\cdot}11Z\ .$$

For other ages the I.Q. y is obtained by subtracting multiples of β, according to equation (6).

Fitting of straight line for $x_0 = 49{\cdot}5$

·350	−·128	·478	11	5·258
·225	−·118	·343	9	3·087
·166	−·060	·226	7	1·582
·156	·110	·046	5	·230
·126	·098	·028	3	·084
−·073	·090	−·163	1	−·163
·950	−·008	·958		10·078

Slope = −10·078/286 = −·03524.

Values of Z and I.Q.s for age 10 : 8

x	Z	x	Z	I.Q.* for 10 : 8 (y_1)
		95		(133·00)
89·5	1·589	90	1·621	127·39
		85		(122·82)
79·5	1·045	80	1·068	119·04
		75		(115·97)
69·5	·674	70	·691	113·34
		65		(110·84)
59·5	·356	60	·371	108·51
		55		(106·35)
49·5	·080	50	·093	104·31
		45		(102·35)
39·5	−·179	40	−·166	100·39
		35		(98·34)
29·5	−·459	30	−·444	96·19
		25		(94·03)
19·5	−·772	20	−·752	91·54
		15		(88·37)
9·5	−1·243	10	−1·215	84·54
		5		(80·00)

* Figures in brackets are interpolated or extrapolated.

In the conversion table I.Q.s have been given correct to the nearest whole number. Various checks may be made on the accuracy of the table : we may note, for example, that the median score of 46·4 and median age 11 : 1½ correspond to an I.Q. of 100, and that in each month-group the percentages of children with I.Q.s of over 100 and 115 (say) are near the expected values of 50 per cent. and 16 per cent. respectively.

REFERENCES

1. Thomson, G. H. (1932). 'The standardization of group tests and the scatter of intelligence quotients.' *Brit. Journ. Educ. Psychol.*, II, 92 and 125.
2. Pearson, K. (1930). *Tables for Statisticians and Biometricians*, Part I, 3rd ed., 1, Table I.

LINEAR AND NON-LINEAR DISCRIMINATING FUNCTIONS

By A. LUBIN

University College and Maudsley Hospital, London

I. *Problem.* II. *Mathematical Derivations.* III. *The Prediction of Qualitative Attributes from Qualitative Categories.* IV. *A Non-parametric Test of Significance.* V. *Summary.*

I. PROBLEM

The main purpose of this paper [1] is to present a summary of recent work in discriminant functions. Discriminant functions represent a fairly new addition to the statistical techniques that can be used by psychologists. Primarily they are solutions to problems in prediction. Given certain measures of physical or mental traits, can we predict that occupation in which an individual will succeed ? Can we state that therapy is most likely to benefit patients who have a certain pattern of scores on a personality test ? In these cases we wish to predict a qualitative attribute of an individual from known quantitative variables. The discriminant function is used whenever there are a set of mutually exclusive classifications and an individual must be assigned to one of these classes on the basis of a standard set of quantitative scores.

Historically speaking, the discriminant function was used long before it was given a name. In 1936, R. A. Fisher described a discriminant function as a linear combination of a set of quantitative variables, which is used to assign an individual to one of two mutually exclusive groups (11). But psychologists had used what Burt has called a 'multiple point biserial' correlation for some time (3, 4, and refs.). The point biserial procedure was derived from the assumption that all individuals within one group should have the same score on the criterion. There was a widespread belief that such a procedure was illegitimate. If psychological measurements are sampled from the normal distribution, the assumption of a point distribution becomes untenable. But definitive justification of the 'multiple point biserial' correlation appeared in the literature when it was shown that the regression weights given by the multiple biserial (which assumed a normal distribution of the criterion), the multiple point biserial (which assumed a point distribution), and the discriminant function of R. A. Fisher were all mutually proportional (4, 5, 30).

Properly to appreciate the importance of the discriminant function we must place it in its setting among other prediction procedures. Table I illustrates four possible types of simple prediction problems.[2] In the last line the problem is raised of how to predict from one set of qualitative variates to another. Although there are special solutions for this problem in terms of quantifying the qualities, the general solution has not yet been stated in purely qualitative terms. One solution is considered in Section III, pp. 99–100.

---

[1] As will be seen, the writer is indebted to Professor Sir Cyril Burt for many of the ideas in the article. Acknowledgement must also be made to Dr. M. Hamilton, Mr. A. Jonckheere, Mr. A. Summerfield, Miss D. Webb, and my colleagues at University College and Maudsley Hospital who have been invaluable sources of stimulation and instruction.

[2] Somewhat similar classifications will be found in 4 and 16 : cf. also this *Journal*, I, pp. 139–40.

TABLE I

Dependent Variate	Independent Variate(s)	Method of Solution
1. Quantitative	Quantitative	Multiple Regression
2. Quantitative	Qualitative	Analysis of Variance
3. Qualitative	Quantitative	Discriminant Function
4. Qualitative	Qualitative	?

Probably most users of analysis of variance have never thought of it as a method of predicting a quantitative variable from one or more qualitative variables. However, the simple probability models used in setting up the equations of analysis of variance postulate a quantitative effect for each category of classification. They can, therefore, be regarded as prediction equations.

Placing discriminant functions in the same table as analysis of variance and multiple regression implies that the discriminant technique is of the same order of importance. In the writer's opinion this is true. The field of vocational and educational guidance alone should find the discriminant function an indispensable tool. In all fields of empirical enquiry, there will be a shift of emphasis. We will no longer ask only the question, " Are the differences between the groups statistically significant ? " The new questions will be, " Are these differences of any practical use ? Can they be used to allocate individuals to their proper classification ? "

II. MATHEMATICAL DERIVATIONS

The Prediction of Quantitative Variates from Qualitative Attributes. The simplest situation in analysis of variance is the one-way classification. As a prediction problem, it is a question whether a single quantitative attribute, such as intelligence, can be predicted from a single qualitative attribute, such as membership in a political party. The basic equation is :

$$x_{ij} = A_j + z_{ij} , \qquad (1)$$

i.e., the observed score, x , of the ith individual in the jth class is equal to the overall level of the jth class (A_j) plus some positive or negative error effect. Thus A_j is the best estimate of the score that can be made from knowing that the individual comes from the jth class. The error of prediction is z_{ij}. The least square estimate of A_j will obviously be the mean of the n_j individuals in the jth class (which will be denoted by $\bar{x}_{.j}$). We will therefore assign to each member of the class j the score $\bar{x}_{.j}$.

The attribute of class membership has in effect been quantified by equation (1). What was a series of classes can now be represented by a series of points on a continuum represented by the x variable. It is this metricization of the qualitative attribute of class membership that makes the linear discriminant function possible. We shall later derive the linear discriminant function from this type of reasoning. It will be important to remember that equation (1) is essentially an estimation function. It gives a predicted score ($x_{.j}$) for each individual which minimizes the sum of squared discrepancies.

The Pearson product-moment correlation between the predicted score $\bar{x}_{.j}$ and the observed score x_{ij} will equal the ratio of the standard deviation of the predicted score to the standard deviation of the observed score. This correlation is the correlation ratio, *eta*. And we may recall that giving the mean of each class as the predicted score for that class maximizes the correlation ratio. (See 16, pp. 264–8.)

We shall now give the usual analysis of variance equations leading up to the F ratio. These equations will be used later in the derivation of the linear discriminant function.

Substituting the least-square estimate for A_j we can write from (1) that $z_{ij} = x_{ij} - \bar{x}_{.j}$. Subtracting the general mean $\bar{x}_{..}$ from both sides of equation (1) and substituting for z_{ij}, we get the standard form

$$(x_{ij} - \bar{x}_{..}) = (\bar{x}_{.j} - \bar{x}_{..}) + (x_{ij} - \bar{x}_{.j}) \qquad (2)$$

The deviation of the observed score from the general mean equals the deviation of the group mean from the general mean plus the deviation of the observed score from the group mean. If we square each term for each individual and sum over the n_j individuals in each group and then over the c groups, we find the well-known important result that

$$\sum_{j=1}^{c}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_{..})^2 = \sum_{j=1}^{c} n_j(\bar{x}_{.j} - \bar{x}_{..})^2 + \sum_{j=1}^{c}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_{.j})^2 \qquad (3)$$

The deviance about the general mean equals the deviance between group means plus the deviance within groups. The equation for the square of the correlation ratio can be written as

$$R^2 = \frac{\text{deviance between groups}}{\text{total deviance}} . \qquad (4)$$

The equation for the F ratio can be written as a function of R^2

$$F = \frac{N - c}{c - 1} \cdot \frac{R^2}{1 - R^2}, \qquad (5)$$

which shows the basic relation between the two statistics.

Derivation of the Linear Discriminant Functions as Canonical Variates. We can now use the above five equations to metricize qualities and derive a linear function which will best predict the quantized attributes. The derivation has been given by R. A. Fisher for the discrimination of two groups (11). An extension to more than two groups was suggested by Burt for psychological and educational purposes (4); and (with a different line of approach) has also been described in an article by C. R. Rao (23). Rao's somewhat brief derivation was given in terms of maximizing Mahalanobis's D^2. Since the D^2 statistic is unfamiliar to most psychologists, the derivation here (which is similar to that of Burt) will be in terms of maximizing R^2, which will also maximize the F ratio.

The F ratio (or its equivalent, the correlation ratio) tells us if class membership significantly differentiates the individual with respect to the quantitative variable. We will now invert the problem of predicting quality from quantity, and assign numbers to each category from the quantitative variables in such a way as to maximize the F ratio.

The question is put in this manner : Given q test scores for the ith person in the jth class, how can we form the linear combination of the q scores that makes R^2 a maximum ?

Let y be the best linear function and let $u_1, u_2, \ldots u_q$ be the best q weights. The notation will be : t represents the tests and runs from 1 to q ; j represents the groups and runs from 1 to c ; i represents the individuals and runs from 1_j to n_j within each group. $N = \sum_{j=1}^{c} n_j$. Then the y score for each individual will be the sum of all his x scores with each x weighted by the appropriate u.

$$y_{ij} = u_1x_{1ij} + u_2x_{2ij} + \ldots u_tx_{tij} + \ldots u_qx_{qij} = \sum_{t=1}^{q} u_tx_{tij} . \qquad (6)$$

In matrix notation :

$$y_N = X_{ij}u ,$$

where y_N is a column vector of N elements, one for each person ; X_{ij} is the raw score matrix of N individuals for q tests, whose general element is x_{tij} ; and u is the column vector of weights, one for each of the q tests, which will specify the best linear discriminant function. N is assumed to be greater than q.

Let us substitute for y_N in the analysis of variance equations and state the formula for the squared correlation ratio, R^2, in terms of X_{ij}, u, and matrices derived from X_{ij}. Thus $(y_{ij} - \bar{y}_{..}) = (y_{ij} - \bar{y}_{.j}) + (\bar{y}_{.j} - \bar{y}_{..})$ in matrix notation becomes

$$Xu = X_i u + X_j u \,, \tag{7}$$

where X is an N by q matrix whose general element is $(x_{tij} - \bar{x}_{t..})$, the deviation of the observed score on test t from the general mean.

X_i is an N by q matrix whose general element is $(x_{tij} - \bar{x}_{t.j})$, the deviation of the observed score on test t from the mean of the jth group.

X_j is an N by q matrix whose general element is $(\bar{x}_{t.j} - \bar{x}_{t..})$, the deviation of the group mean from the general mean.

u is the set of weights defining the linear discriminating function. In equation (3) the sum of squares or deviance about the general mean was equal to the deviance between group means plus the deviance within groups. Writing the same equation for y in matrix notation:

$$u'X'Xu = u'X_i'X_i u + u'X_j'X_j u \,. \tag{8}$$

Our notation will be simplified if we use the following definitions.

1. $X'X \equiv G$, the "general" dispersion or codeviance matrix. The general element in the sth row and the tth column is $g_{st} \equiv \sum_{j=1}^{c} \sum_{i=1}^{n_j} (x_{tij} - \bar{x}_{t..})(x_{sij} - \bar{x}_{s..})$. It is a square symmetric matrix of order q, as are the two matrices defined below, W and B. The element $g_{st} = N r_{st}\, \sigma_s\, \sigma_t$.

2. $X_i'X_i \equiv W$, the "within groups" codeviance matrix.

$$w_{st} = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (x_{sij} - \bar{x}_{s.j})(x_{tij} - \bar{x}_{t.j}).$$

3. $X_j'X_j \equiv B$, the "between groups" codeviance matrix.

$$b_{st} = \sum_{j=1}^{c} n_j\, (\bar{x}_{s.j} - \bar{x}_{s..})(\bar{x}_{t.j} - \bar{x}_{t..}).$$

These three q by q codeviance matrices are analogous to the single variable deviances in equation (3). For example $G = W + B$.

The formula for the square of the correlation ratio of the discriminating function, y, is similar to equation (4)

$$R^2 = u'Bu / u'Gu \,. \tag{9}$$

The formula for R^2 will now be used to derive the linear discriminating function. We shall seek the vector u which makes R^2 a maximum. The procedure will be to differentiate R^2 with respect to u, set the derivative equal to zero and solve the resulting equation for u. We obtain

$$\frac{dR^2}{du} = \frac{(u'Gu)\, 2u'B - (u'Bu)\, 2u'G}{(u'Gu)^2} = 0. \tag{10}$$

Multiply equation (10) by the scalar $(u'Gu)$, substitute $R^2 = u'Bu/u'Gu$ from equation (9), simplify, and the result is

$$u'B - R^2\, u'G = u'\,(B - R^2 G) = 0. \tag{11}$$

Post-multiplying by G^{-1}

$$u'\,(BG^{-1} - R^2 I) = 0. \tag{12}$$

This, in slightly different notation, is the equation reached by Burt (4). As he has suggested, it can also be derived direct from the ordinary expression for a canonical correlation, and can readily be extended to the multidimensional case by means of a preliminary factor analysis (4; 6, p. 105).

Equation (12), and the conditions of the problem, tell us that the u we seek is the latent vector corresponding to the largest latent root of BG^{-1}. This can be found by the usual iterative methods (4). BG^{-1} is non-symmetric, and one must be careful to *pre-multiply* by the trial row vector. The transposed matrix $G^{-1}B$ must be used if *post-multiplication* by a trial column vector is desired. The largest latent root is the largest R^2 obtainable.

Equation (12) also implies that there are as many latent roots which are solutions as there are linearly independent rows or columns of $B\,G^{-1}$. The number of independent linear discriminating functions is therefore equal to the rank of $B\,G^{-1}$. There are definite limits to the rank of $B\,G^{-1}$. G must be a non-singular matrix, i.e., must be of rank q, or else the inverse G^{-1} will not exist. On the other hand, B has a maximum rank of $c - 1$. This is a result of the fact that the sum of any column of X_j equals a constant, zero. This means that the maximum number of independent rows in X_j is $c - 1$. So if q is greater than $c - 1$, the rank of $B\,G^{-1}$ is $c - 1$. If $c - 1$ is greater than q, the rank of $B\,G^{-1}$ is q.

While obviously the best single linear discriminating function is that one which is associated with the largest latent root, the set of all linear discriminating functions defines the number of dimensions which the groups occupy in q dimensional space.

Multivariate linear discriminating analysis can be likened to a form of factor analysis which extracts only those factors which help to distinguish groups from one another. I therefore suggest using the phrase "canonical variates" to denote linear discriminating functions extracted in this manner. Any linear combination of scores on which the group means tend to be equal will be discarded automatically. This raises the possibility of experimental verification of postulated factors. If a factor can be defined in terms of groups of homogeneous individuals, then the discriminating function which maximizes the discrimination between these groups satisfies the definition of the postulated factor.

The set of latent vectors associated with $B\,G^{-1}$ will form a q by r matrix which we can call U. U cannot be compared directly with the usual factor analysis matrix F. F is usually used to denote the set of correlations of q tests with r uncorrelated factors. It is therefore the set of regressions of the tests on the factors. U is the set of regressions of the r discriminating functions on the q tests when the standard deviations of the tests have not been made equal to unity. F can be turned into a matrix of regressions of factors on tests by one of several formulæ. $R^{-1}F$ is one such formula (where R is the q by q correlation matrix). When multiplied by the appropriate diagonal matrix of standard deviations, these factor regression matrices will be directly comparable to U. Tests of significance for hypothetical discriminating functions versus observed discriminants have been discussed by Rao (24) and Fisher (14). But special procedures will have to be worked out for verifying factors found in previous factor analysis studies by comparing U with F.

It is only permissible to define a factor operationally in terms of a linear discriminating function, when that factor has itself been defined in terms of a discrimination between groups. H. J. Eysenck, however, in his studies of the 'dimensions of personality' speaks of factors which are defined in terms of their power to differentiate between various diagnostic types of mental disorder. In a later article I hope to explain my view more fully by means of illustrative examples, using (amongst other material) data drawn from Dr. Eysenck's own work.

Multivariate discriminating analysis can also be used to answer the question of whether a set of categories really form a single linear continuum. If the rank of U is greater than one, then the groups cannot be represented as points or intervals along a single scale. The best linear discriminating function would still not suffice to represent the differences between the groups. For example: discriminating function analysis could be used to test if political groups such as Liberal, Conservative, and Socialist actually fall on a single dimension of Radicalism–Conservatism.

When the covariance matrices of all groups are assumed equal, the best discriminating functions are all linear in form. A full discussion of this case with a worked example is given by Rao and Slater (25). A further discussion of this point will be found in Section II (b) of this paper.

The Prediction of Qualitative Attributes from Quantitative Variables by the Maximum Likelihood Discriminating Function. The underlying weakness in the usual approach to discriminating functions is that unless the categories lie on a single dimension they cannot be adequately discriminated from one another. Quantifying the categories may force an actual multidimensional set of qualities on to a single straight line. The procedure may really be distorting the data and obscuring real differences between the categories.

The maximum likelihood discriminating function developed by Rao avoids this difficulty (23). In this method, the pattern of scores is used to discriminate between groups of individuals without the step of condensing each individual's set of scores into a point on a linear continuum. This basically different approach releases the psychologist from the necessity of forcing quantification on essentially qualitative data.

A full discussion of the mathematical basis for the method will be found in the paper by Rao. More elementary discussions will be found in the papers by Welch (29) and C. A. B. Smith (27).

Smith applied the maximum likelihood method to the problem of discriminating two groups. His solution was a likelihood ratio which, interestingly enough, simplifies to Fisher's linear discriminating function when the variance–covariance matrices of the two groups are equal.

Briefly stated, the method consists of calculating the likelihood for each individual with respect to each group. The individual is then assigned to the group for which he has the maximum likelihood. Rao's full method would lead to as many likelihood scores for each individual as there were groups. Smith's solution was to calculate the likelihood ratio, the ratio of the likelihood of belonging to one group to the likelihood of belonging to the other group. When there are only two groups, this is a single score. It is possible to examine the logical basis of the maximum likelihood solution without going deeply into the mathematics of the problem. Smith gives a simple explanation of the two-group case. In the following we shall consider the simplest case that will illustrate the full non-linear discriminating method, namely, the case of three groups. Let us, however, begin by glancing at the *univariate* or linear case.

Given three groups of people, Group A, Group B, and Group C, who have taken a single test X, how can we allocate each individual to a group, merely from a knowledge of his score on X, so as to minimize the percentage of misclassifications? The logical structure of this question is identical with the problem of quantitative prediction. A set of mutually exclusive groups or categories is a qualitative variable just as a set of scores is a quantitative variable. Allocation to a group is equivalent to prediction of a quantitative score. And the concept of minimizing the percentage of misclassified individuals is of the same importance and plays the same role as the least squares concept of minimizing the squared errors of prediction. (It will be recalled that the canonical linear discriminating function described in Section I was derived from least square considerations.)

A graph of the three groups is given in Fig. 1. x_1, x_2, and x_3 are individuals possessing scores on X at the point indicated. The ordinate L_{c1} represents the number of Group C individuals that may be expected to have the score x_1. Similarly, the ordinates L_{a1} and L_{b1} represent the number in Groups A and B that will have the score x_1.

We cannot take any specific individual who has the score x_1 and say that these frequency elements correspond to the probabilities that he has come from these various groups. Since we have specified the membership of each individual at the beginning of our problem, the probability that any individual belongs to a specific group is either zero or unity.

However, the concept of 'likelihood' as used by R. A. Fisher may be used here. The three ordinates represent the three likelihoods that a person with score x_1 will belong to each of the three groups. Rao has demonstrated the important theorem that if each person is assigned to the group for which his likelihood is greatest, the number of misclassifications is a minimum. In this case all persons having the score x_1 would be assigned to Group A. Persons with scores x_2 and x_3 would be assigned to Group B and Group C respectively. If there were equal numbers in each group it is obvious that this would lead to a minimum number of misclassifications. If the number of people in the universe differs from the sampled number in the group, then each frequency element must be weighted by an amount proportional to the number in the group for this universe. In what follows it will be assumed that the proportions of the three groups in the universe are equal.

It will be noted that there is an interval, a, along the X axis where La is always larger than Lb or Lc. Similarly there is an interval b where the likelihood of belonging to Group B is greater than the other likelihoods. And for all scores in the interval c, the likelihood is always greatest for Group C. Rao's theorems state that such regions exist and can always be found if the ordinate for each score in each group is known.

Now the point which separates the regions is the point where the likelihood functions are equal, i.e., the point where the curves intersect. If we knew the distribution function for Group A, then the number of individuals from Group A who would be misclassified could be calculated. The number misclassified as Group B is equal to the integral of La over the interval b. Similarly the number of Group A individuals misassigned to Group C is the area under the A frequency curve for the interval c. It will be noted that as yet we have not assumed any specific type of distribution curve; and the above statements are true no matter what form the distribution function takes.

As Burt has shown (see Appendix, p. 104), the point X where two normal curves intersect is given by the formula

$$X = \frac{(\sigma_1^2\bar{x}_2 - \sigma_2^2\bar{x}_1) \pm \sigma_1\sigma_2\sqrt{\left(\bar{x}_1 - \bar{x}_2\right)^2 + 2\left(\sigma_1^2 - \sigma_2^2\right)\left[\ln \frac{N_1\sigma_2}{N_2\sigma_1}\right]}}{\sigma_1^2 - \sigma_2^2}$$

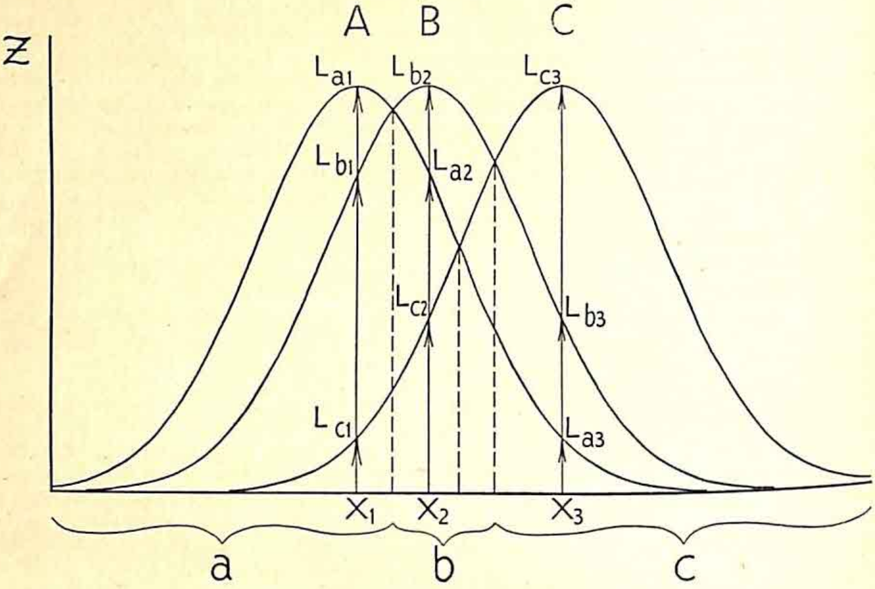

FIG. 1. Discrimination : the one-dimensional problem.

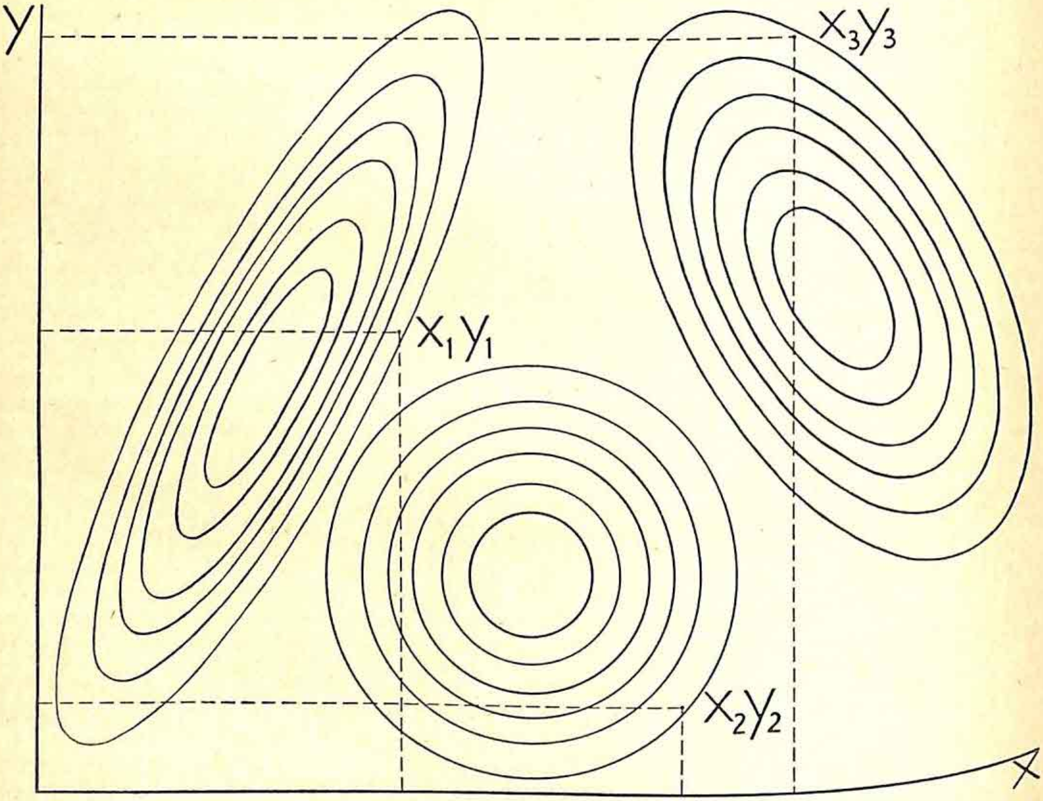

FIG. 2. Discrimination : the two-dimensional problem.

where $X \equiv$ the raw score at the point of intersection, $\bar{x}_1 \equiv$ the mean of the first group, $\bar{x}_2 \equiv$ the mean of the second group, $\sigma_1 \equiv$ the standard deviation of the first group, $\sigma_2 \equiv$ the standard deviation of the second group, $N_1 \equiv$ the number in the first group, $N_2 \equiv$ the number in the second group. When the N's and σ's are equal the intersection point is midway between the two means. Burt, who first obtained the above equation (2), seems to have been the first to propose the point of intersection as the cut-off point for allocation of individuals into groups.[1]

Let us now consider the bivariate case. We can plot each person on a two-dimensional graph with x and y as the axes. If we think of frequency as a third dimension orthogonal to the plane of the graph, then each group resembles a hill. The height of the hill represents the relative frequency with which a particular pair of scores x_1y_1, x_2y_2, etc., is found in the group. Fig. 2 is a rough schematic representation of the situation.

In Fig. 2 contour lines have been used to indicate height, i.e., the frequency element. It is obvious that there are regions where all individuals have a greater likelihood of belonging to one group than the other two. Rao has outlined methods of obtaining these regions (23). For our purposes, the fact that they exist is sufficient.

In this bivariate case it would be necessary to compute the three likelihoods for each pair of scores and assign each person to the group where his likelihood is greatest. The number of misclassifications resulting from this procedure is always equal to or smaller than the misclassifications that would arise from using only one variable.

The same procedure can easily be extended to the case of more than two variables. For each person, the three likelihoods (one for each group) would be calculated for each set of scores. Allocation would be to the group with the highest likelihood. It is clear that increasing the number of groups merely increases the number of likelihoods that must be calculated. The maximum likelihood discriminating technique is, then, applicable to any number of groups. Increasing the number of quantitative variables increases the complexity of the likelihood functions but does not increase the number of likelihoods that must be computed.

A real difficulty arises in the actual calculation of the likelihoods. The number of individuals in any group is, of course, finite. Empirical frequencies could be determined for grouped values of one variable. But the task would soon become impossible if we attempted to determine the relative frequencies of each pair or triplet of scores. To actually determine these proportions empirically would demand an impossibly large number of people in the sample. This same difficulty arises in the prediction of qualitative attributes from other qualitative variables. It is discussed in Section III.

The only practical procedure appears to be an assumption about the theoretical nature of the distribution within each group. Special considerations may lead to the selection of a binomial, a rectangular, a Poisson, or other frequency distribution as best representing the shape of the distribution curve within each group. In many, if not most, cases the normal multivariate function will prove most convenient.

To use a theoretical frequency distribution it will be necessary to calculate, for each group, the constants which define the function. In the case of a normal univariate distribution, the mean and standard deviations are sufficient. The score of any individual may then be substituted into the equation for each group. Therefore each individual will have as many likelihoods as there are groups.

The equation for the multivariate normal distribution will involve the calculation of all means, standard deviations, and inter-correlations. The inverse of each matrix of inter-correlations must be computed. The likelihood function for each person will involve a generalized quadratic function whose coefficients are the elements of the inverted covariance matrix. This process has been described in detail by C. A. B. Smith (27). Since Smith is concerned only with the two-group case, he uses as his discriminant function the ratio of the two likelihoods, or, rather, twice the natural logarithm of the likelihood ratio. It will be recalled that for the multiple-group case it is necessary only to select the largest likelihood. Even so the likelihood function for the multivariate normal distribution is likely to be simplified if logarithms instead of the actual frequencies are used. Another aid to simplification is to reduce the quantitative variables which are being used as predictors to as few orthogonal variables as possible. The canonical reduction connected with the best linear discriminating function (see equation (12)) seems likely to furnish the optimum set of predictors. But this elaborate procedure is unworkable for a large number of variables. Some simplified form of *a priori* factor analysis such as that suggested by L. S. Penrose (20) might prove convenient and relatively efficient. The condensed set of variables should be so selected as to be linearly independent since the multivariate normal distribution requires the inverse of the correlation matrix. This inverse will not exist if the variables are not linearly independent of one another.

---

[1] *Mental and Scholastic Tests* (1921), p. 165. The conditions of the problem assume that we choose the point of intersection *between* the two means, i.e., take the sign before the radical to be positive. In certain exceptional cases the relative size of σ_1 and σ_2 or of N_1 and N_2 may make this impossible. Note that, here and below, I use "ln" as an abbreviation for "natural logarithm of," i.e., for "$\log_e$."

(a) *The Non-Linear Discriminating Function for the Case of a Multivariate Normal Distribution.* The multivariate normal distribution is of special interest because so many psychological variables are distributed normally. When the variables used are themselves the sum of a large number of variables, it is very likely that these sums are distributed approximately as the normal curve. This follows from the central limit theorem (9, p. 232). It is possible in the psychological field to construct variables that closely approximate a normal distribution of quantitative scores by obtaining the sum of a large number of repeated measures of the same trait. The discriminating equation will accordingly be derived assuming that the distribution for each category or group is multivariate normal.

A matrix notation will conveniently be used. To distinguish vectors from scalars, the letters denoting the variable elements will be underlined when they represent vectors. Let $C_j \equiv$ the $q \times q$ covariance matrix of the jth group, $\underline{x}_i \equiv$ the 1 by q row vector of observed raw scores for the ith individual, $\underline{d}_{ij} \equiv \underline{x}_i - \underline{m}_j \equiv$ the 1 by $\overline{q}$ row vector of the q scores expressed as deviations from the q means of the jth group; $\underline{d}_{ik} = \underline{x}_i - \underline{m}_k$, the same raw score vector expressed as a deviation from the vector of means for the kth group. And let $n_j \equiv$ the number of individuals in the jth group; $N \equiv \sum_{j=1}^{c} n_j$; $p_j \equiv n_j/N$; and $f_j(\underline{x}_i) \equiv$ the frequency of the set of observed raw scores $\underline{x}_i$ in group j.

There will be c discriminating functions, one for each group. The standard operating procedure would be to calculate $f_j(\underline{x}_i)$ for all the c groups and then assign all individuals having the pattern of scores, $\underline{x}_i$, to the group where the frequency of $\underline{x}_i$ is highest. This will minimize the total number of misclassifications.

The equation for the frequency of $\underline{x}_i$, assuming a multivariate normal distribution, is:

$$f_j(\underline{x}_i) = \frac{n_j}{(2\pi)^{q/2} \mid C_j \mid^{\frac{1}{2}}} \quad \exp\left[-\tfrac{1}{2}\, \underline{d}_{ij}\, C_j^{-1}\, \underline{d}_{ij}'\right]. \tag{13}$$

Similarly the frequency for $\underline{x}_i$ in the kth group would be:

$$f_k(\underline{x}_i) = \frac{n_k}{(2\pi)^{q/2} \mid C_k \mid^{\frac{1}{2}}} \quad \exp\left[-\tfrac{1}{2}\, \underline{d}_{ik}\, C_k^{-1} \underline{d}_{ik}'\right]. \tag{14}$$

Thus the frequency of any pattern of scores, $\underline{x}_i$, can be calculated for each of the c groups. We can simplify equation (13) somewhat. The term $(2\pi)^{q/2}$ is a constant for all calculated frequencies for all observed values of $\underline{x}_i$, and does not need to be considered.

Let $$L_{ij} \equiv +\,2 \ln\left[f_j(\underline{x}_i)\right] + q \ln 2\pi. \tag{15}$$

Then $$L_{ij} = 2 \ln(n_j) - \ln \mid C_j \mid - \underline{d}_{ij}\, C_j^{-1} \underline{d}_{ij}'. \tag{16}$$

L_{ij} is a quadratic expression which can be used in place of the frequency. When the frequency of $\underline{x}_i$ is near zero, L_{ij} will tend to minus infinity. When the frequency of $\underline{x}_i$ is unity or greater than unity, L_{ij} will be positive.

There is still another form of the discriminating function that is useful when the absolute number of individuals in each group is not known, but the relative proportions are given.

Let $$\lambda_{ij} \equiv 2 \ln\left[\frac{f_j(\underline{x}_i)}{N}\right] + q \ln 2\pi. \tag{17}$$

Then $$\lambda_{ij} \equiv 2 \ln p_j - \ln \mid C_j \mid - \underline{d}_{ij}\, C_j^{-1} \underline{d}_{ij}'. \tag{18}$$

Where $$p_j \equiv n_j/N. \tag{19}$$

λ_{ij} will vary from minus infinity to the maximum positive quantity of $(q \ln 2\pi)$.

Thus a personnel officer may not know what total number of bombardiers, navigators, and pilots will be required. If, however, he knows what the relative proportions should be in the Air Force as a whole, he can use λ_{ij}. This will allocate all available individuals to air crew training schools with the smallest number of misclassifications.

We have thus shown that, when the distribution within each group is multivariate normal, the best discriminating formula for each group is a quadratic expression of the q scores for each individual.

(b) *The Linear Discriminating Function for the Case of a Multivariate Normal Distribution.*

Let us assume that $$C_j = C_k = \ldots = C. \tag{20}$$

Then $$\lambda_{ij} = 2 \ln p_j - \ln \mid C \mid - (\underline{x}_i - \underline{m}_j)\, C^{-1} (\underline{x}_i - \underline{m}_j)'. \tag{21}$$

The quantity $(-\ln \mid C \mid)$ will be constant for all groups.

$$\lambda_{ij} + \ln \mid C \mid = 2 \ln p_j - \underline{x}_i C^{-1} \underline{x}_i' + 2\underline{x}_i C^{-1} \underline{m}_j' - \underline{m}_j C^{-1} \underline{m}_j' \,. \quad (22)$$

In equation (22) it can be seen that the quantity $-(\underline{x}_i C^{-1} \underline{x}_i')$ is a constant and does not vary over the c groups. It is now possible to define a new term, l_{ij}, which will be directly related to the likelihood of $\underline{x}_i$ in group j; but will involve only linear functions of $\underline{x}_i$

$$l_{ij} = \tfrac{1}{2}(\lambda_{ij} + \ln \mid C \mid + \underline{x}_i C^{-1} \underline{x}_i') = \underline{x}_i C^{-1} \underline{m}_j' + \ln p_j + \tfrac{1}{2}\underline{m}_j C^{-1} \underline{m}_j' \,. \quad (23)$$

Equation (23) states that l_{ij} is calculated for any $\underline{x}_i$ by applying the q weights given by $C^{-1} \underline{m}_j'$ and adding the constant term, $\ln p_j - \tfrac{1}{2} \underline{m}_j C^{-1} \underline{m}_j'$. When there are only two groups, l_{ij} can be shown to be directly proportional to the point-biserial multiple regression function, or to Fisher's linear discriminating function. Rao and Slater give a worked example for three variables and six groups (25) in which they use a function called L_i. In this article $l_{ij} \equiv L_i$.

Summary. Section II (a) shows that for a multivariate normal distribution in each group, and unequal covariance matrices, a quadratic function of the observed scores is the best discriminating function. Two such quadratic functions have been derived, L_{ij} and λ_{ij}, where

$$L_{ij} = 2 \ln [f_j (\underline{x}_i)] + q \ln 2\pi = 2 \ln n_j - \ln \mid C_j \mid - \underline{d}_{ij} C_j^{-1} \underline{d}_{ij}' \quad (24)$$

and

$$\lambda_{ij} = 2 \ln \left[\frac{f_j (\underline{x}_i)}{N} \right] + q \ln 2\pi = 2 \ln p_j - \ln \mid C_j \mid - \underline{d}_{ij} C_j^{-1} \underline{d}_{ij}' \,. \quad (25)$$

L_{ij} is to be used where the number of subjects, n_j, to be allocated to each of the c groups is known. λ_{ij} is for use when only the percentage of subjects, p_j, to be allocated to each of the c groups is known.

Section II (b) shows that, for a multivariate normal distribution in each group and a constant covariance matrix, C, a linear function l_{ij}, of the observed scores, is the best discriminating function.

$$2\; l_{ij} = 2 \ln \left[\frac{f_j \underline{x}_i}{N} \right] + q \ln 2\pi + \ln \mid C \mid + \underline{x}_i C^{-1} \underline{x}_i'$$
$$\text{or } l_{ij} = \underline{x}_i C^{-1} \underline{m}_j' + \ln p_j - \tfrac{1}{2}\underline{m}_j C^{-1} \underline{m}_j' \,. \quad (26)$$

A consideration of the equations for L_{ij}, λ_{ij}, and l_{ij} will show at once that class membership, as defined in Section II, is not at all an invariant relationship. Whether a recruit is classified as a bombardier or a pilot, by a selection officer, is likely to vary as the quotas for these jobs vary. If an invariant criterion of class membership is wanted, it will be necessary to set a confidence level for rejection from each group. L_{ij} or the other two equations can then be solved for the critical value which represents this confidence level. All individuals falling below this critical value will be rejected from the group.

Recently, there has been a great deal of interest and research on the problem of using test scores in differential diagnosis of mental disorders. Rapaport (26), Jastak (18), Rabin (21), and many others have advocated the use of certain difference scores and/or ratios of one score to another to allocate patients to various psychiatric categories. In particular, Jastak has concluded that ratio scores are most appropriate.

Obviously, this is the very type of problem for which the discriminating function was invented. From our previous discussions it should be clear that whenever the distribution of scores is multivariate normal (or can be satisfactorily approximated by a multivariate normal distribution) no ratio score can be an optimal method of discrimination. A difference score can be effective only if the variances and covariances are constant over the groups. The optimal discrimination will be achieved by using a quadratic function of the scores.

III. THE PREDICTION OF QUALITATIVE ATTRIBUTES FROM QUALITATIVE CATEGORIES

In Table I above, where various types of prediction problems were tabulated, the question of predicting qualities from other qualities was stated to have no known solution. Actually many statistical techniques have been devised which deal with contingency tables, the basic data for this type of prediction problem. But all of these techniques test the deviation from a chance distribution, or measure the amount of association in a manner comparable to the correlation coefficient. What is needed is a workable solution to the question of how to predict class membership from knowledge of membership in a set of other classes.

Suppose a population of N individuals is divided into categories $A_1, A_2, \ldots A_j, \ldots A_p$ which we wish to predict. These will be called the criterion categories. Suppose we are given the further information that each person has a secondary classification in categories $B_1, B_2, \ldots B_i, \ldots B_q$ which are to be used to predict the criterion categories. These will be called the predictor categories.

We can now make out a $q \times p$ matrix whose general element is n_{ij}, the number of people who are in both the row B_i and the column A_j. The principle of maximum likelihood gives a direct simple solution for the best prediction of A from B. For any specified row B_i, find the n which is the largest in the row of values, $n_{i1}, n_{i2}, \ldots n_{ij}, \ldots n_{ip}$. If n_{i1} is the largest frequency in the B_i row, then the best prediction from B_i is to A_1. Thus a particular A category will be predicted from each B category. More than one B category will sometimes predict the same A category. There is no necessity for each B category to be linked with a different A category. It is easy to see that this method of taking the A category which has the highest frequency in a particular row and making it the prediction for that B row results in a set of predictions that minimizes the total number of misclassifications. This is the simple situation that prevails in the case of the single predictor variable.

In the case of two predictor variables, we are given the additional information that the individual has a tertiary classification on categories $C_1, C_2, \ldots C_h, \ldots C_r$. This leads to a three-dimensional table of $q \times p \times r$, whose general element is n_{hij}, the number of people who have a joint membership in C_h, B_i, and the criterion category A_j. The maximum likelihood method for the case of two predictors is formally the same as the case of prediction from B alone. For each combination C_h, B_i, find the prediction category A_j which has the largest frequency. When this category A_j is taken to be the prediction for an observed joint membership in C_h and B_i, again a minimum number of misclassifications will result.

But the number of possible combinations of B and C is qr. As the number of discriminators is increased arithmetically, the number of combinations increases geometrically. Since any sample is finite, as the number of predictors is increased, a point will soon be reached where the predicted A category from a particular subclass of the discriminators will be indeterminate. There will be no individuals who fall into this particular subclass.

Another difficulty is that the predictions must be established by a detailed investigation of empirical frequencies in each case. There seems to be no method whereby a set of parameters can be calculated which, when applied to the observed set of predictors, will give the best prediction of A.

Each contingency table could be first tested by the usual chi-square technique to ensure that the observed frequencies are unlikely to have arisen by chance. Coefficients of association can be computed for contingency tables. But the best measure, in view of the method we are using, is the percentage of incorrectly classified individuals.

R. A. Fisher has presented an interesting technique in his *Statistical Methods for Research Workers* (15) for quantizing the qualitative variates in a two-way $q \times p$ contingency table. Scores are assigned to categories A and B in such a way that the Pearson product-moment correlation between the quantized variates is a maximum. The extension to the multivariate case does not seem obvious. The basic difficulty with this approach is that it assumes exactly what we wish to avoid—that each category is a point on a linear continuum.

On the whole, the solutions presented here seem inadequate to the writer. Perhaps the presentation of the problem in this *Journal* will focus attention on the question and reveal that some existing technique can be used.[1] Most responses to situations are easier to classify in a qualitative rather than a quantitative manner; and the psychologist would find a great deal of use for a technique that required no intervening quantification.

---

[1] Professor Burt suggests treating each composite class as a ' determinate ' for a new ' determinable,' and then using the method of maximum probability, if the values are independent, and, if not, the usual weighting formula ($w'R^{-1}$) : see *Factor Analysis of Qualitative Variables*. In theory the standard probability calculus can always be used. Possibly a model could be devised which allowed a ' multivariate multinomial ' distribution to be employed, deduced from its first and second moments.

IV. A NON-PARAMETRIC TEST OF SIGNIFICANCE

All discriminating functions, no matter whether they are based on a mathematical equation or a subjective judgment, can be summarized in a two-way contingency table of observed categories A (in which the individual is observed to fall) against predicted category B. Using the notation of Section III, the columns A_1 to A_p will represent the observed categories. Rows B_1 to B_p will be the predicted classes such that if discrimination were perfect everyone in B_1 would be observed to fall in category A_1, everyone placed in B_2 would fall in A_2, etc. For perfect discrimination the square contingency table would have the diagonal cells filled, but the off-diagonal entries would be zero.

Let the frequency in the ith row and the jth column be denoted by n_{ij}. Let the marginal total of the ith row be B_i and the marginal total of the jth column be A_j. Since B_i is in principle adjustable, we can always, theoretically, make B_i equal to A_i. In this discussion we shall not make that simplification. We shall treat the general case when A_i and B_i are not necessarily equal.

We can now define a pure chance discrimination where correct classifications occur by chance alone. For a pure chance discrimination

$$n_i = \frac{A_i B_j}{N}, \text{ where } N = \sum_{i=1}^{p} \sum_{j=1}^{p} n_{ij}. \tag{27}$$

Our real interest is in the diagonals of the square contingency table. The misclassifications in the off-diagonal cells may be occurring in a manner contrary to a hypothesis of chance distribution. But this is immaterial except for theoretical analysis of the mistakes. The point is whether the correct classifications are occurring more frequently than chance would allow. A chi-square test would therefore not be exactly appropriate. A contingency table in which the diagonal entries were significantly greater than chance expectation would necessarily have a significant chi-square. But a significant chi-square is not sufficient to show that the discriminating function has made a greater number of correct classifications than would be expected by chance. However, the chi-square test should always be made and if it is not significant the discriminating function on the whole does not do better than chance.

If we are interested in the total number of correct classifications over all observed groups, then we must test the difference between the two quantities. Let

$o \equiv \sum_{i=1}^{p} n_{ii} \equiv$ number of persons observed to be correctly classified ;

$e \equiv \sum_{i=1}^{p} \frac{A_i B_i}{N} \equiv$ theoretical number of correct classifications expected by chance.

Then, the null hypothesis will be that the number of individuals who are correctly classified does not differ significantly from the number who would be correctly classified if the assignment to the categories were based on chance alone : that is, if B_1 individuals were picked at random from the total group of N individuals, the number of such B_1 individuals who will also be in category A_1 will be $\frac{A_1 B_1}{N}$.

In cases where the number of correctly classified individuals, o, is less than or equal to the chance number e, it is not necessary to test for significance. We are only interested in testing those cases where o is greater than e. This is essentially a one-tail test. We are sampling from a population where o can take all values from zero to N. But we are only interested in the portion of the curve from e to N. For small values of N the binomial $\left\{\frac{e}{N} + \frac{N-e}{N}\right\}^N$ should be evaluated. The test of significance would then be to sum all terms of the binomial, from $\frac{N!}{o!\,(N-o)!}\left\{\frac{e}{N}\right\}^o\left\{\frac{N-e}{N}\right\}^{(N-o)}$ to the last term $\left\{\frac{e}{N}\right\}^N$. This sum would give the probability that the number of correct classifications obtained by chance would reach or exceed the number o, on the assumption that e denotes the chance expectation in the general population.[1] When N is large, the binomial distribution is approximately symmetric. If the probability distribution is symmetric, the usual t ratio test for the difference between two percentages can be applied, viz., $t = \dfrac{(o - e)\sqrt{N}}{\sqrt{e\,(N - e)}}$.

---

[1] For a discussion of this procedure in psychological work and references dealing with the mathematical basis, see Burt, *The Backward Child*, App. III, 'The hypergeometric formula from double sample,' where convenient methods of calculation are also described.

The 99 per cent. confidence interval and the 95 per cent. confidence interval of the binomial distribution have been presented by C. J. Clopper and E. S. Pearson (7). Tables of these values are given by Snedecor (28, pages 4–5).

Even though the observed total number of correct classifications, o, may be significantly greater than chance, it may be suspected that not all predictions are being made with the same accuracy. The n_{ii} for each B_i can be tested against the $e_i \left(\equiv \dfrac{A_i B_i}{N}\right)$ in the same manner as the overall test of o and e. Here the base number would be the number in the category B_i, rather than N.

The usual t ratio test for the difference between two proportions can be made if it is suspected that category B_i is a better discriminator than B_j. However, note that n_{ii} must be divided by B_i and n_{jj} by B_j to give the correct proportions for the test.

In this way, fairly precise tests may be made of the effectiveness of any discriminating function without having to make any assumptions as to normality of distribution, equal covariance matrices, etc. Johnson (19) gives a practical procedure for determining when the normal curve may be expected to represent the binomial : (1) compute $9N/(N + 9)$; (2) if $e > 9N/(N + 9)$ use the normal curve.

Occasionally we may wish to compare the efficiency of two different discriminating functions on the same data. This leads us to a two by two table. The two rows would represent the individuals (1) who have been correctly classified, and (2) who have been incorrectly classified by the first discriminating function. The two columns would similarly split the groups into correct and incorrect classifications made by the second discriminating function. The four cells would represent (*a*) those individuals correctly classified by both functions ; (*b*) those individuals correctly classified by the first function, but incorrectly by the second ; (*c*) those individuals correctly classified by the first discriminant and correctly by the second ; (*d*) those individuals incorrectly classified by both functions.

With this fourfold table in mind, we may now ask whether p_1, the proportion of individuals correctly classified by the first discriminant, is greater than p_2, the proportion of correct classifications by the second discriminant. For a large enough number of cases, an appropriate test of significance would be

$$\frac{(p_1 - p_2)\,N}{\sqrt{p_1 q_1 + p_2 q_2 - 2r_{12}(p_1 q_1 p_2 q_2)^{\frac{1}{2}}}} = c\,,$$

where N is the total number of individuals to be classified and r_{12} is the point fourfold correlation coefficient. When N is large enough, c will be normally distributed and may be evaluated at the usual levels of significance tabled for the normal curve.

V. SUMMARY

This paper has considered the problems of predicting qualitative attributes from (1) quantitative variables and (2) qualitative variables. The canonical variates which define the minimum subspace within which a set of group means may lie have been derived from the basic equations used in analysis of variance.

The maximum likelihood discriminating functions for the prediction of qualitative attributes from quantitative variables have been described and explained. Explicit formulæ have been given for the case of a multivariate normal distribution with (1) unequal covariance matrices and (2) equal covariance matrices.

A maximum likelihood solution is proposed for the problem of predicting qualitative attributes from qualitative variables.

Some non-parametric tests of significance have been suggested which are applicable to all discriminating functions.

REFERENCES

1. Brown, G. W. (1947). ‘Discriminant functions.’ *Annals of Math. Stat.*, XVIII, 514.
2. Burt, C. (1921). *Mental and Scholastic Tests.* London: King.
3. Burt, C. (1938). ‘The unit hierarchy.’ *Psychometrika*, III, 151.
4. Burt, C. (1938). ‘The discriminant function’ (unpublished Laboratory Notes).
5. Burt, C. (1944). ‘Statistical problems in the evaluation of army tests.’ Psychometrika, IX, 233.
6. Burt, C. (1948). ‘Factor analysis and canonical correlations.’ *Brit. J. Psych., Stat. Sect.*, I, 105.
7. Clopper, C. J., and Pearson, E. S. (1934). ‘The use of confidence or fiducial limits illustrated in the case of the binomial.” *Biometrika*, XXVI, 404.
8. Cochran, W. G., and Bliss, C. I. (1948). ‘Discriminant functions with covariance.’ *Annals of Math. Stat.*, XIX, 151.
9. Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton: Princeton University Press.
10. Cronbach, L. J. (1949). ‘Pattern tabulation.’ *Educ. Psychol. Meas.*, IX, 149.
11. Fisher, R. A. (1936). ‘Use of multiple measurements in taxonomic problems.’ *Ann. Eugen.*, VII, 179.
12. Fisher, R. A. (1938). ‘The statistical utilisation of multiple measurements.’ *Ann. Eugen.*, VIII, 376.
13. Fisher, R. A. (1939). ‘The sampling distribution of some statistics obtained from non-linear equations.’ *Ann. Eugen.*, IX, 238.
14. Fisher, R. A. (1940). ‘The precision of the discriminant function.’ *Ann. Eugen.*, X, 422.
15. Fisher, R. A. (1941). *Statistical Methods for Research Workers.* (8th ed.) Edinburgh: Oliver and Boyd.
16. Guttman, L. (1941). ‘An outline of the statistical theory of prediction,’ 253–312, in *The Prediction of Personal Adjustment*, edited by Paul Horst. Ann Arbor, Michigan: Edwards Brothers.
17. Hotelling, H. (1931). ‘The generalisation of student’s ratio.’ *Ann. Math. Stat.*, II, 360.
18. Jastak, J. (1949). ‘Problems of psychometric scatter analysis.’ *Psychol. Bull.*, XLVI, 177.
19. Johnson, P. O. (1949). *Statistical Methods in Research.* New York: Prentice-Hall, Inc.
20. Penrose, L. S. (1947). ‘Some notes on discrimination.’ *Ann. Eugen.*, XIII, 228.
21. Rabin, A. I. (1941). ‘Test score patterns in schizophrenia and non-psychotic states.” *J. Psychol.*, XII, 91.
22. Rao, C. R. (1946). ‘Tests with discriminant functions in multivariate analysis.’ *Sankhyā*, VII, 407.
23. Rao, C. R. (1948a). ‘Utilisation of multiple measurements in problems of biological classifications.’ *Roy. Stat. Soc. Jour. (Series B)*, X, 159.
24. Rao, C. R. (1948b). ‘Tests of significance in multivariate analysis.’ *Biometrika*, XXXV, 58.
25. Rao, C. R., and Slater, P. (1949). ‘Multivariate analysis applied to differences between neurotic groups.’ *Brit. J. Psych., Stat. Sect.*, II, 17.
26. Rapaport, D. (1945). *Diagnostic Psychological Testing*, Vol. I. Chicago: Year Book Publishers.
27. Smith, C. A. B. (1947). ‘Some examples of discrimination.’ *Ann. Eugen.*, XIII, 272.
28. Snedecor, G. W. (1946). *Statistical Methods* (4th ed.). Iowa: Iowa State Collegiate Press.
29. Welch, B. L. (1939). ‘Note on discriminant functions.’ *Biometrika*, XXXI, 218.
30. Wherry, R. J. (1947). ‘Multiple bi-serial and multiple point bi-serial correlation.’ *Psychometrika*, XII, 189.

APPENDIX

On the Discrimination between Members of Two Groups (Note by C. Burt).[1] Let there be two groups whose means ($\bar{x}_1$ and $\bar{x}_2$), standard deviations (σ_1 and σ_2), and numbers (N_1 and N_2) are known, and whose frequency-distributions are normal but overlap. It is required to find a point on the abscissa, x_i say, which will best serve to discriminate between members of the two groups. To fix our ideas, we may suppose that the first and larger group (denoted by a subscript 1) consists of all the pupils in the ordinary elementary schools, and that the second (denoted by a subscript 2) consists of all those who have been diagnosed as mentally defective by the school medical officers, and transferred to special schools: as tested by the re-standardized Binet scale in 1921, the distribution of the two groups was approximately normal. Our object will then be to find a borderline, in terms of an I.Q., which will minimize the amount of misclassification.

The number of children misclassified will be represented by the sum of two areas, roughly triangular, namely, the upper tail of the left-hand curve added to the lower tail of the right-hand curve. Thus the sum to be minimized will be

$$\int_{x_i}^{+\infty} \frac{N_1}{\sigma_1\sqrt{2\pi}}\, e^{-\frac{(x-\bar{x}_1)^2}{2\sigma_1^2}}\, dx + \int_{-\infty}^{x_i} \frac{N_2}{\sigma_2\sqrt{2\pi}}\, e^{-\frac{(x-\bar{x}_2)^2}{2\sigma_2^2}}\, dx = \int_{x_i}^{+\infty} \varphi_1(x)\, dx + \int_{-\infty}^{x_i} \varphi_2(x)\, dx = I_1 + I_2, \text{ say.}$$

To find the value of x which will minimize this sum, we must differentiate it with respect to x, and set the first derivative equal to zero. The first derivative will be

$$\frac{dI_1}{dx} + \frac{dI_2}{dx}.$$

By the usual rules for differentiating a definite integral with respect to a lower or an upper limit, this expression $= -\varphi_1(x_i) + \varphi_2(x_i)$. On setting this equal to zero, we obtain $\varphi_1(x_i) = \varphi_2(x_i)$: that is,

$$\frac{N_1}{\sigma_1\sqrt{2\pi}}\, e^{-\frac{(x_i-\bar{x}_1)^2}{2\sigma_1^2}} = \frac{N_2}{\sigma_2\sqrt{2\pi}}\, e^{-\frac{(x_i-\bar{x}_2)^2}{2\sigma_2^2}}.$$

But these expressions also represent the ordinates of the curves 1 and 2 at the point x_i. We have thus shown that the required value for x_i is the value that marks the point on the base-line between the two means, *at which the ordinate has the same height for both curves;* and this ordinate in turn marks the point where the two curves intersect. Hence, as might have been guessed on intuitive grounds, the best way of discriminating between the two distributions is to cut between the two overlapping curves at the point where the trough between them is lowest.

To express x_i as an explicit numerical function of the constants describing the two curves, we have merely to take logarithms and solve for x_i. We thus obtain the formula given in the text (p. 95 above, and *Mental and Scholastic Tests*, 1921, p. 165). It was on this basis that an I.Q. of 70 was suggested as the borderline for the mentally defective. It should be noted that the proof assumes that we already know the proportion of N_1 to N_2; and the known values must be used, not—as is so often done—either equal numbers or the arbitrary numbers contained in the samples selected by testing. Since the two triangular tails are not exactly equal, the application of the formula will *not* in practice yield *exactly* the required numbers N_1 and N_2. If the psychologist's instructions require him to fill exactly N_2 vacant places in the special schools, then another procedure would have to be used, viz., selecting the dullest N_1 on the I.Q. scale (see *loc. cit.*, p. 167). Further, if his business is to find the best-weighted battery of tests to discriminate two groups, classified on the basis of what is really a continuous, quantitative, and approximately normal variable (such as 'intelligence' or 'neuroticism'), then a discriminant function derived from a criterion that is qualitative only in form is likely to lose much of the information that might have been gained by using a graded criterion.

---

[1] The following extract from (4) has been appended at Mr. Lubin's request.

THE INFLUENCE OF DIFFERENTIAL WEIGHTING

By CYRIL BURT

I. The Problem of Weighting. II. The Effects of Different Types of Weighting. III. Weighted Sums as Factors. IV. Experimental Studies of Differential Weighting. V. Summary and Conclusions.

I. THE PROBLEM OF WEIGHTING

Weighting in Psychology. In considering how to combine the marks obtained from an examination that includes several question-papers, or from a test battery comprising a number of subtests, one of the most puzzling problems is that of differential weighting. Is anything to be gained by assigning different weights to the component tests or papers, and if so what are the best methods to adopt? In spite of their apparent importance, surprisingly little attention has been given to the issues involved. What Freeman wrote in 1926 is even truer to-day: "Weighting has come to be far less commonly used than it was a number of years ago"[1]; and most practical psychologists who comment on the matter are content to quote Guilford's statement that "weighting is not worth the trouble"—a statement, I fancy, which was intended to apply not so much to the tests in a battery as to the items in a test.[2]

In their book on *The Scientific Study of Educational Problems*—one of the few later publications that touch upon the wider problem—Monroe and Engelhart have briefly reviewed the relevant literature. They point out that the majority of American writers have been concerned chiefly with the influence of weighting on *reliability*, and that the conclusion generally drawn is that "weighting does not affect the reliability to any marked degree."[3] On the other hand, as they rightly say, "the effect of weighting on the *validity* of the measures has received relatively little attention" (11, p. 190). Accordingly, before turning to the statistical questions that appear to be involved, it may perhaps help those who are more interested in practical methods than in technical discussions if I begin by describing in some detail the situations in which such problems have come to the fore.

(a) *Practical Importance*. The earliest instances are to be found in the field of ordinary examinations. For example, in the elementary school system teachers were accustomed to promote pupils year by year on the basis of a composite mark obtained in terminal examinations in all the subjects of the curriculum ; certain pupils were not infrequently recommended for a backward class or for possible certification as mentally defective on the ground of low marks ; other pupils were awarded scholarships on the results of a composite examination in a prescribed set of subjects. Discussions continually arose as to the relative weight to be attached to the different subjects : should not Arithmetic be given a higher weight than English, and should not both receive a far higher weight than Manual work ? In such cases, it is to be noted, the interest centres on the single composite mark, intended to represent the 'highest common denominator' or 'factor' ; and this factor here relates to a more or less closed and restricted field—namely, the fixed group of subjects forming the elementary syllabus.

---

[1] *Mental Tests*, 1926, p. 275. In support of his own conclusions he quotes Douglass, H. R., and Spencer, P. L., 'Is it Necessary to Weight Exercises in Standard Tests?', *J. Educ. Psych.*, XIV, 1923, pp. 109–12. For later discussions, coming to much the same views, see Corey, S. M., 'The Effect of Weighting Exercises in a New Type of Examination,' *J. Educ. Psych.*, XXI, 1930, pp. 383–5, Potthoff, E. F., and Barrett, N. E., 'A Comparison of Marks, Based on Weighted and Unweighted Items,' *ibid.*, XXIII, 1932, pp. 92–8, Scates, D. E., and Noffsinger, F. R., 'Factors which Determine the Effectiveness of Weighting,' *ibid.*, XXIV, 1933, pp. 280–5.

[2] *Psychometric Methods*, 1936, p. 448.

[3] This no doubt arises from the facts that, in the theory of least squares, weighting is discussed chiefly in connection with the allowances to be made for differences in reliability (2, 9), and that, in the theory of psychophysical measurement, it is again the differences in the reliability of the observations that require some form of weighting to be introduced (cf. Guilford, *loc. cit.*, pp. 54, 176, and refs.). But, as I pointed out in the memorandum cited below, the problems with which the practical psychologist is concerned depend far more on validity, and thus resemble those arising in the computation of composite 'index numbers' in economics (a point also made by Edgeworth: cf. 1 and 8).

Similar questions arose for discussion during the investigations of the *International Institute Examinations Enquiry*. In a memorandum which I was asked to prepare, an attempt was made to derive certain general formulæ which would indicate the effects of equal and of differential weighting on reliability and validity ; a factorial procedure was suggested ; and arithmetical illustrations were given comparing the results obtained by different methods.[1] The main conclusion emerging was that the chief danger lay, not so much in the neglect of carefully calculated differential weights, but in the unrealized presence of random or unintentional weighting—particularly that due to accidental differences in the standard deviations of the examiners' marks. In the published report, to save space and excessive technicality, the formulæ and their algebraic derivations were not reproduced in full ; and both the inferences drawn and the practical recommendations made have since elicited some criticism. More recently, in reporting an investigation carried out for the Scottish Council for Research in Education, and after fully discussing 'The Weighting of Measures,' Dr. McClelland and his colleagues conclude by saying that, "in view of our results, we find it impossible to lay down a single principle, even of the most general character" (14, p. 83). And finally the whole problem has cropped up afresh in connexion with the weighting of tests used in the fighting services during the war and the weighting of examination marks and other assessments for the allocation of pupils to different types of secondary schools.

(b) *Theoretical Importance*. From a theoretical standpoint the question has a still wider interest. Without a clear appreciation of the effects of differential weighting, the nature of the various modes of factor analysis can hardly be understood. But in theoretical discussions on factor analysis the factors envisaged are usually taken as defined, not by denotation (as something common to a definite and circumscribed curriculum of subjects or performances), but rather by connotation (as a quality distinctive of a certain field or domain, of which the processes actually tested or measured form merely specimens or representative samples). Nevertheless, from a statistical standpoint, "a factor is still nothing but a weighted average of a set of measurements empirically obtained, being weighted according to the degree in which it demonstrably represents the field in question"[2]; and, if this description were kept in mind, the common confusion between factors as statistical components and factors as concrete entities (such as 'abilities,' 'propensities,' and the like)—a confusion against which Professor Thomson has so rightly and so often protested—would less frequently occur.

There is a further confusion which has not so frequently been noted, namely, between the different types of statistical component, regardless of any psychological interpretation. Thus, any factor which has positive saturations for all examples belonging to the same broad category, field, or genus is constantly described as a 'general factor,' regardless of the particular weighting procedure by which the so-called general factor has been extracted : in the field of cognitive processes, for instance, the 'highest common factor' in a bipolar analysis, the 'basic factor' in a group factor analysis, the first 'principal component' in a Pearson analysis, and 'Spearman's *g*' are all treated as if they referred to precisely the same fundamental concept, even by those who would scrupulously refrain from identifying that concept with 'general intelligence.'[1]

---

[1] Burt, C., 'The Analysis of Examination Marks,' *ap.* Hartog, P., and Rhodes, E. C., *The Marks of Examiners*, 1936, p. 303, cf. pp. 245f. As will be seen from the memoranda, Dr. Rhodes was at first highly doubtful, on quite general grounds, about the validity of factorial procedures, and proposed instead to calculate the 'ideal mark' by "taking a weighted average, the weights being proportional to the degrees of precision, which are inversely proportional to the squares of the standard deviations" (p. 193). Subsequently, however, he reached very much the same formulæ as myself (cf. p. 319) ; and in a later publication of great interest showed how factorial methods might usefully be applied in at least one non-psychological field (Rhodes, E. C., 'The Construction of an Index of Business Activity,' *J. Roy. Stat. Soc.*, C, 1937, pp. 18f.).

[2] Cf. Burt, C., *Factors of the Mind*, p. 74. Spearman, it is true, regularly objected that, as he interpreted the term, a factor was not, and could not be, either an average or a weighted sum (cf. *Abilities of Man*, pp. 61–5). I have, however, already attempted to meet his objections (*loc. cit.*, pp. 211f.), and need not discuss them further here. I may add that, when I first introduced such phrases as 'highest common factor' (3, 4) and suggested ways of calculating it, I had in mind a 'factor' in an abstract mathematical sense. However, now that the term is used so frequently to designate a concrete quality, it might be better (as I have suggested elsewhere) to adopt the more accurate term 'component' for the factor *before* identification, and reserve the term 'factor' for the ability or other concrete quality with which that component is eventually identified.

Although in this paper I shall treat the numerical constants that we are trying to determine as no more than weighting coefficients in an equation of estimation, I do not mean to deny that some of them may conceivably have a theoretical significance of their own. Just as the factor-measurements, if they could be appropriately selected and accurately calculated, would prove to be important constants for describing a given individual, so, I believe, the factor-saturations obtained with a suitable choice of tests would often turn out to be important constants describing certain fundamental types of mental activity. In quantitative psychology, the enquiries which lead to a factor analysis mark only the initial stages of our work. Having by that means satisfactorily classified and defined the chief modes of independent variation, we must next press on to determine, wherever we can, those functional equations which are required for laying the foundations of a sound theory of experimental psychodynamics. In such a theory what we now call factor-saturations will, I imagine, play a new role, similar to that of specific heats and dielectric constants in physics or the valencies of the elements in chemistry. This seems confirmed by several recent studies on mental work. But I here refer to it solely to remind the reader that I am far from regarding factor analysis and the calculation of weighting coefficients as the final goal of quantitative research.

Partial Regression Equations. In what follows, we shall be concerned merely with the problem of combining measurements, which are assumed from the outset to form acceptable (though doubtless somewhat fallible) indications of some specified dispositional property—some generalized characteristic which we desire to measure for practical purposes, and which for convenience can be called a factor. Whether such a factor exists and what is its nature are questions we may suppose to have been already decided. The school psychologist, for example, may have to accept the fact that within the schools concerned there is only one line of promotion, and consequently that the promotion of this pupil or of that will depend on their marks for what I called a factor of 'general educational ability'[2]; and his immediate problem will be simply to decide what are the relative weights to be attached to papers or tests in English, Arithmetic, Reading, Spelling, Handwork, and the like.

In the earliest factorial studies the main object was to procure individual measurements for the broader factor assumed to underlie all forms of cognition—'general intelligence,' as it was commonly termed. As soon as it was realized that a quality like 'intelligence' could best be estimated, not by applying a single test (as Spearman and others originally assumed), but by combining a wide variety of tests (as Galton, Binet, and their followers had maintained), the question at once arose as to whether these different tests should be differently weighted, in accordance with their difficulty, their nature, their reliability, or their validity, and, if so, how the weighting coefficients should be computed. Thus, Yerkes, in his 'Point Scale' revision of the Binet series, contended that we ought to prescribe different maxima for the different test-problems, and put forward various *a priori* considerations for their determination.[3] My own contention was that, in principle, "each test should be weighted, not equally, nor yet arbitrarily, but according to an empirical *regression coefficient* based on its ascertained *correlation* with intelligence itself."[4]

---

[1] This type of confusion is analogous to that which old-fashioned logicians used to point out between a 'generic *differentia*' and a 'generic *proprium*' (a *proprium* is a 'property' in the last of the four traditional senses recognized by the scholastic writers, namely, *id quod pertinet omni et soli et semper*). Thus, to take the stock scholastic instance, if we define man as a 'rational animal,' rationality is his '*highest* common factor'; 'ability to laugh' is merely *a* common or 'general' factor, not the most essential or fundamental.

[2] He is, of course, often tempted to ask how far such a promotion scheme can really be justified; and in that case one further advantage of the statistical treatment suggested is that it enables us to offer a definite reply to this question also. Thus, up to the age of 10, the single factor of *general* educational ability appears to account for nearly 50 per cent. of the individual differences in the achievements of the various subjects of the curriculum; whereas each of the remaining factors (*special* educational abilities) accounts for only 10 per cent. or less. On the other hand, at the age of 12 the general factor accounts for little over 25 per cent.: (cf. 6, p. 266; also *Brit. J. Educ. Psych.*, IX, p. 55, XIII, p. 132, and refs.).

[3] Yerkes, R. M., *et al.*, *A Point Scale for Measuring Mental Ability*, 1915. For criticisms of the weighting proposed, see *Mental and Scholastic Tests*, p. 72. Cf. also Binet, A., 'On the Necessity of Establishing a Scientific Diagnosis of Inferior Intelligence,' *L'Année Psych.*, XI, 1905, pp. 163f.

[4] Burt, C., *Mental and Scholastic Tests*, 1921, p. 74. The first suggestion that regression coefficients should be used in this way was contained in my article on 'The Measurement of Intelligence by the Binet-Simon Tests,' *Eug. Rev.*, VI, 1914, pp. 140–52.

In support of this suggestion, I argued that " if we adopt Professor Pearson's device of 'multiple correlation,' it can easily be demonstrated that a far better assessment can be achieved with a representative sample of tests, covering a number of different cognitive processes, each appropriately weighted, than can be obtained with only a single test, or even a mere total of marks from several tests taken as they stand." In seeking in this way to secure more trustworthy measures for the 'general factor' of intelligence, we are in effect taking samples or specimens of different types of intelligent performance, much as, in endeavouring to procure a trustworthy estimate of (say) the proportional amount of water in human muscle, the physiologist examines specimens of all the chief types of muscular tissue. The only novelty is that we attempt to weight each specimen according to its representative value.

External and Internal Criteria. By what criterion, however, are we to determine the correlation of each test with the quality it is intended to measure? In the case of general intelligence, it was at first assumed that the independent estimates supplied by teachers or other observers would provide a valid standard. On that assumption the diagnostic value of each specimen test could be assessed by comparing or correlating its results with this external criterion, and appropriate weights could then be computed by the principle of least squares, in such a way that the resultant mark should yield the closest approximation to the estimates which the criterion supplied. Unfortunately, in many investigations no such external standard was available, and in others the independent estimates provided were not sufficiently trustworthy to be used for that purpose. The method of factor analysis was accordingly developed, specifically to meet cases of this kind. But essentially the same formulæ still seemed applicable for calculating the weights, although the criterion was now based on internal rather than on external evidence.

Substitutes for Regression Equations. The full procedure, however, though plausible enough in theory, proved decidedly laborious in practice. It entailed lengthy calculations, first to determine the regressions or weights and then to compute the weighted sums. Accordingly, for practical purposes various simplified methods were commonly substituted, without any very full justification being offered for their use.

(i) *Unweighted Sum of Standardized Measurements.* In the early days of factorial work, the most usual procedure was to reduce all the test-measurements to a comparable unit, either by using standard measure, or by converting the test-measurements into ranks, and then to sum or average the measurements so standardized. Tests which had a low correlation with all the other tests, and therefore with the general factor, were generally rejected. Thus, in my 1909 article on 'Experimental Tests of General Intelligence' (3), out of the whole series of twelve tests, I retained the best six, and averaged the ranks for those six. This was found to raise the correlation from 0·76 (the coefficient furnished by the best single test) to 0·85.

(ii) *Saturations as Weights.* In later work, where more accurate factor-measurements were required (e.g., for the supplementary factors), the saturation coefficients themselves were not infrequently used as approximate weights. But sometimes a test with a positive saturation may have a negative regression (see Appendix, p. 124).

(iii) *Simplified Weights.* More often a rough and ready compromise between the two procedures was adopted. Instead of saturation coefficients, which entailed multiplication by two- or three-figure decimal fractions, simple unit-figure weights (e.g., 1, 2, 3) were adopted for the better tests, and the poorer tests omitted as before (a more recent example of this procedure will be found in the Scottish investigation on selection for secondary education, 14, pp. 82f.). This kind of weighting can be most easily introduced by allowing 1, 2, or 3 marks for each correct answer according to the weight to be given to the several tests in a carefully selected battery, or by prescribing as the maximum mark for certain tests or question papers a figure two or three times as large as that for the others. In compiling booklets of tests, a common plan was to increase or reduce the number of problems in each test in such a way that the marks obtained for each test at once eliminated the effects of a varying standard deviation, and gave the better tests a higher loading, without any explicit multiplication being required when computing the totals.

Whatever method of weighting is adopted, it should be remembered that the real effects of a set of weights must be assessed in terms of the *relative contribution of each weighted component to the total variance* ; and this will depend, not only (i) on the assigned or ostensible weights, but also (ii) on the standard deviation of each component, and (iii) on its correlation with each of the other components (see Appendix, p. 124).

Weighting in a Non-Psychological Research. It is interesting to note that, in attempting to determine the relative ' crop productivity ' of different British counties, Professor Kendall (13) has recently used two alternative devices, not unlike those I have described. They are, as he says, " based on psychological work, and measure a *factor* which may reasonably be regarded as covariant with productivity by regarding crop yields as analogous to test scores." His first method is to calculate what I should call ' factor-measurements ' for the ' highest common factor,' i.e., for the ' general factor of productivity ' ; and for this purpose he employs a partial regression equation derived in accordance with Pearson's method of ' principal axes.' His second method is a simplified procedure similar to the first of those mentioned above, namely, simply ranking the counties for their production in each of the main types of agricultural produce, and then taking the average rank to indicate general productivity. He finds that " the closeness of the results, which are reached by widely different methods, is surprising " : the correlations between the two sets of figures, obtained for four different years, all fall between 0·95 and 0·97 (inclusive). Thus, so far as these examples go, " the average rank tends to give the same ranking as one type of general factor loading." He remarks that he has " been unable to find any mathematical proof that a close correspondence will always exist " ; and he asks " whether a similar problem has been studied in psychology " (13, pp. 25, 45).

II. THE EFFECTS OF DIFFERENT TYPES OF WEIGHTING

Special Problems. In examining the general effects of different modes of weighting, the problems involved can, I fancy, be most usefully studied by considering the process in four stages, namely, (*a*) *a priori* or subjective weighting, (*b*) random weighting, (*c*) equal weighting, and (*d*) *a posteriori* or *ad hoc* weighting (i.e., weighting based on empirical data analysed specially for the purposes of the particular research).

(*a*) A Priori *Weighting*. In the early days of mental and scholastic measurement,[1] the methods of weighting proposed were usually based on subjective considerations. The highest weights were assigned to those aspects of the subject which were deemed, on *a priori* grounds, to be of greatest moment, or to those measurements which were presumed to be the most reliable. Such principles are still often followed in the field of education. Examiners, for instance, frequently allot different maxima to different question-papers according to the supposed importance of the subjects ; and, as a rule, no very clear distinction is drawn between ' weighting for value ' and ' weighting for validity.'[2] *A priori* weighting also seems needed where the criterion appears to be composite in content, i.e., to be influenced by several components, which may be independent, alternative, or in some cases even compensatory (as in occupational success).

---

[1] The earliest discussion of the need for a systematic determination of weights in attempts at mental measurement is to be found in Edgeworth's papers on ' The Statistics of Examinations ' and kindred topics : *J. Roy. Stat. Soc.*, LI, 1888, pp. 599–635 ; LIII, 1890, pp. 460–75, 644–63. Cf. also Galton, *Brit. Ass. Ann. Rep.*, 1889, pp. 471f.

[2] I here use the term ' validity ' to include ' precision ' or ' reliability ' on the ground that diminished reliability is one cause of diminished validity. By ' weighting for value ' I mean attempts in which the weights proposed are based on judgments of practical utility or cultural worth rather than on diagnostic validity or reliability. Thus, at a meeting of examiners for the junior county scholarship examination, certain members urged that the maximum allotted to the English paper should be higher than that allotted to the Arithmetic paper, " because in a civilized English-speaking community it is more important for its members to write English than to do sums." The validity of each paper as an index of innate general ability was not discussed. For an interesting suggestion as to how " weighting for social importance " might be carried out, see Kelley, T. L., *Essential Traits of Mental Life*, 1935, pp. 70, 92–5 : especially equations (iii) and (47).

In this respect the mode of weighting was not unlike that employed in calculating composite 'index numbers'[1] by early economists; and thus open to very similar criticisms (cf. 7, p. 212, for a remark which applies admirably to the psychologist's computations). *A priori* weighting is occasionally used in other scientific fields. But, according to those most competent to give a judgment on the matter, "'empirical' weighting, that is, weighting according to individual judgment" (or, as the psychologist would prefer to term it, 'subjective' or 'impressionistic' weighting), "is always a matter of difficulty, and at best its results must remain open to objection."[2] Accordingly, I venture to suggest that even here a factorial procedure might prove suggestive.

Even where those concerned are unanimously agreed that trait x (say) has a higher social or cultural value than trait y, and therefore deserves a weight which is n times as large as that assigned to y ($n > 1$), there seems to be no generally recognized principle for deciding what numerical value shall be assigned to n. Accordingly, I suggest what may be called the principle of virtual frequency. To allot to x a weight which is n times that allotted to y, means that we are in effect treating each of the x-measurements as if they had been based on *n times as many independent observations* as the y-measurements.

I may add, however, that, alike in psychological testing and in school or university examinations, wherever I have been able to check the results of such subjective weighting against objective criteria, its value has proved extremely doubtful. I should thus readily endorse the view expressed by Sir Percy Nunn. In a memorandum on 'Methods of Marking' (dealing primarily with the Teachers' Certificate Examination), he has stated that, "with rare exceptions, no satisfactory case can be made for allotting different weights to different examination papers merely on personal or *a priori* grounds. The wide differences between examiners as to what the weights should be, and the erratic effects on mark-lists, suffice to demonstrate that in most cases subjective weighting is just random weighting. What the precise effect of such random weighting may be it is scarcely possible to predict. But one thing can be said with certainty: its general tendency must be to reduce, seldom to enhance, the accuracy of the final marks."[3]

(*b*) *Random Weighting*. As I have already noted, it is too often forgotten that, when a set of measurements or marks are summed or averaged just as they stand, a certain amount of weighting is unintentionally, and often quite unconsciously, introduced, owing to the fact that the raw marks or measurements have different standard deviations. Such differences in the scatter are likely to have no relation either to the value of the separate tests, or to any common factor underlying the whole set of tests or to any valid criterion which the tests or marks are intended to estimate. Thus their effects are equivalent to that of random weighting. Since, then, both deliberate weighting and unintentional weighting so often tend to introduce a large element of random weighting, it seems desirable to determine, if we possibly can, what (in Nunn's words) "the precise effect of such random weighting" is likely to be.

Let us, therefore, begin by examining the effect of weights, selected absolutely at haphazard, on the weighted sums so obtained. Let us suppose that we have n sets of assessments, $x_1, x_2, \ldots x_n$, each set measuring the same attribute, x, with different degrees of accuracy, or, it may be, measuring some different but typical manifestation of the same attribute. (If the reader wishes to keep a concrete instance in mind, he may think of the assessments as the result of applying n tests of intelligence to N school children.) Our ultimate aim is to combine these n measures into a single composite measure by means of some appropriate linear function. Let the set of weights for one such linear function be designated U_i, and for

[1] It was for this reason that I ventured to suggest that the methods of factor analysis might be applied to this and other problems in economics (10, p. 313). Bowley's definition of index numbers (5, p. 196) would serve admirably as a definition of factors. The chief difference (which curiously enough is not brought out explicitly by his definition) arises from the fact that index numbers are commonly used to measure variations which are related to time, and hence are expressed as percentages of an arbitrary standard instead of as deviations about the mean. I may add that index numbers, like factors, are of two distinguishable types: (*a*) they may refer to a domain defined connotatively (as with Jevons' monetary index); (*b*) they may refer to a limited domain defined denotatively (as with the index for cost of living): (cf. p. 106 above).

[2] Freund, I., *The Study of Chemical Composition*, 1904, p. 85. Cf. Kohlrausch, *Physical Measurements*, 1894, especially chapter on 'Errors of Observation,' and Ostwald, *Physico-Chemical Measurements*, 1902, chap. I, 'Calculation.'

[3] This opinion was repeated in a later memorandum submitted by Sir Percy Nunn to the International Institute Examinations enquiry.

another W_i $(i = 1, \ldots n)$. Then the two alternative composite measures which we propose to compare may be written

$$y = U_1x_1 + U_2x_2 + \ldots + U_nx_n\,, \tag{i}$$

$$z = W_1x_1 + W_2x_2 + \ldots + W_nx_n\,. \tag{ii}$$

We shall assume that x_i, y, and z are each to be in unitary standard measure.

Now let $\bar{r}$ denote the average of the $\frac{1}{2}n(n-1)$ inter-correlations, $\overline{U}$ and $\overline{W}$ the means, and σ_u^2 and σ_w^2 the variances of the weights. It is then easy to show (see Appendix) that, if the weights have been chosen at random, the correlation between the two composite measures will be approximately

$$r_{yz} = 1 - \frac{(1-\bar{r})}{n\,\bar{r}} \cdot \tfrac{1}{2}\left(\frac{\sigma_u^2}{\overline{U}^2} + \frac{\sigma_w^2}{\overline{W}^2}\right), \tag{iii}$$

$$\text{or } 1 - \frac{(1-\bar{r})}{n\,\bar{r}} \cdot \frac{\sigma_w^2}{\overline{W}^2}\,, \tag{iv}$$

when the means and standard deviations of the weights are the same.

These two expressions indicate very clearly what will be the general effect of random weighting in various actual cases. The following conclusions may be drawn.

(1) The larger the number of tests employed (i.e., the higher the value of n), the higher will be the correlation between the two weighted sums. As the number of tests is increased indefinitely, so the resulting correlation, r_{yz}, tends to become perfect, even though the two sets of weights have been chosen absolutely at random.

(2) The larger the average inter-correlation between the several tests (i.e., the higher the value of $\bar{r}$), the higher will be the correlation between the two weighted sums. As the average inter-correlation approaches 1·00, so the resulting correlation, r_{yz}, tends to become perfect, even though the weights have been chosen absolutely at random. Conversely, the lower the average inter-correlations, the lower will be the value of r_{yz}. An important corollary follows. If, in order to be sure of assessing a factor from different angles, we have employed a highly heterogeneous set of tests, which have low inter-correlations, and if for practical reasons we can only include a small number of tests in our battery, then the choice of appropriate sets of differential weighting becomes of special importance.

(3) The smaller the standard deviations of the weights, the higher will be the resulting correlation between the two weighted sums. If in both sets the standard deviation is zero (i.e., if within each set the weights are identical, though the weights in one set may not be identical with those in the other), then the correlation will be unity. Conversely, with random weights, the larger the standard deviations, the lower will be the correlation. This confirms the familiar observation that, provided the individual weights do not vary widely, the agreement between the two weighted sums tends to be pretty close, whereas when the range of the weights varies widely, the agreement may be greatly reduced.

(4) If in both sets the mean weight is large in comparison with the standard deviation, then once again the agreement between the two weighted sums will be close. It is thus the relative size of the differences between the weights, not their absolute size, that affects the resulting correlation.

(5) Finally, for the correlation between the weighted sums to be positive, both the mean weights must have the same sign. In particular, if in each set the vast majority of the weights or all the positive weights are positive throughout, then the resulting correlation r_{yz} will be positive. This has a special bearing on the calculation of measurements for the 'first factor' in factorial procedures. Whether it is a 'general' factor, a 'basic' factor, or a first 'principal component,' it is in most psychological enquiries likely to have positive saturations for all the tests or traits in the battery (with the possible exception of one or two variables which may act as 'suppressors'). Hence, unless the number of tests or traits is small, gross inaccuracies in the weighting coefficients will impair the resulting measurements far less than might be supposed.

(c) *Equal Weighting*. Let us now turn to the case in which the n standardized tests are given equal weights. This is of special interest on both practical and theoretical grounds. From a practical standpoint it provides by far the simplest procedure for combining a set of measurements into one composite measurement; from a theoretical it is the method of combination implicitly adopted in the analysis of variance (see 15, Table III).

Let us consider first the common situation in which the n tests have been administered twice to the same group. For example, in the 1909 enquiry the six tests mentioned above were applied first by myself and secondly by Dr. Flugel. A straightforward deduction from Pearson's product-moment formula yields the following result :[1]

$$r_{\bar{x}\bar{x}'} = \frac{\Sigma(x_1 + x_2 + \ldots + x_n)\,(x_1' + x_2' + \ldots + x_n')}{\sqrt{\Sigma(x_1 + x_2 + \ldots + x_n)^2}\ \sqrt{\Sigma(x_1' + x_2' + \ldots + x_n')^2}}$$
$$= \frac{\Sigma r_{ii'} + \Sigma\Sigma r_{ij}}{n + \Sigma\Sigma r_{ij}} = \frac{\bar{r}_{ii'} + (n-1)\bar{r}_{ij}}{1 + (n-1)\bar{r}_{ij}} \qquad \text{(v)}$$

where the dash denotes the second application of the test, and $r_{ii'}$ the reliability coefficients (it is assumed that r_{ij} and $r_{i'j'}$ have been pooled to determine what is here called r_{ij} ; $i \neq j$). In the example cited the formula yields $r_{\bar{x}\bar{x}'} = \dfrac{0{\cdot}682 + 5 \times 0{\cdot}423}{1{\cdot}000 + 5 \times 0{\cdot}423} = 0{\cdot}898$. On calculating the correlation direct from the averaged ranks, the value reached was 0·902. (The slight difference is due to rounding.)

If we know the correlation of each test with the criterion (*c*, say), then the correlation of the unweighted sum with the criterion will be

$$r_{\bar{x}c} = \frac{\Sigma r_{ic}}{\sqrt{\Sigma\Sigma r_{ij}}} \qquad (i, j = 1, \ldots n)$$
$$= \frac{\bar{r}_{ic}\sqrt{n}}{\sqrt{1 + \bar{r}_{ij}\,(n-1)}} \qquad (i \neq j)\ . \quad \text{(vi)}$$

By solving for *n* in one or other of these equations it is possible to estimate roughly how many subtests will be required to yield a correlation of a required amount.[2] If the criterion is an internal criterion, based on simple summation, the expression thus reached simplifies to the square of the expression reached in equation (v), with $r_{ii'}$ now representing communalities rather than reliabilities. The reader can easily deduce for himself the corresponding formulæ for an internal criterion derived by weighted summation or any other procedure.

The same argument shows that the correlation of each test (say, the *k*th) with the average of all the tests is

$$r_{kg} = \frac{\Sigma r_{kj}}{\sqrt{\Sigma\Sigma r_{ij}}} \qquad (i, j = 1, \ldots n)\ , \quad \text{(vii)}$$

where r_{ii}, the self-correlation, is here taken to be 1·00. Now, with the majority of correlation tables obtained in psychological testing, an application of this formula would at once reveal that the tests employed have a widely different correlation with the 'general factor,' estimated provisionally as the unweighted average of all the tests. And this naturally suggests that it is a mistake to give all the tests precisely the same weight. Hence we are led to enquire how the best set of differential weights can be procured.

(*d*) *The Best Weights*. If we ignore the possible existence of specific factors, the most obvious method would be to seek the 'line of closest fit.' This was the procedure proposed by Karl Pearson, who defined 'closest' in terms of least squares.[3] He further pointed out that the requisite weights could be conveniently obtained by seeking the "largest principal axis of the correlation type ellipsoid." And it is, as Kendall has observed, "a remarkable fact that all the information we need, even to represent the variation in a linear subspace of *m* dimensions, is contained in the correlation matrix." [4]

---

[1] The principles from which this and kindred formulæ were derived were first developed by Karl Pearson. To psychologists equation (v) is best known from its application by Brown and Spearman to the problem of increasing the reliability of tests by increasing their length (cf. Brown, W., and Thomson, G., *Essentials of Mental Measurement*, 1925, p. 132 and refs.). Brown points out that his formulæ were "based on demonstrations given in lectures by Professor Pearson" (the source from which many of my own were also derived).

[2] See 10, p. 303, where several numerical illustrations are given. It should be emphasized that, as the assumptions involved are never likely to be precisely fulfilled, the indications must be regarded only as approximate.

[3] Pearson, K., 'On Lines and Planes of Closest Fit,' *Phil. Mag.*, II, 7th ser., 1901, pp. 559f.

[4] Cf.17, § 9. I should myself prefer to say "in the table of covariances and variances." This includes the correlation matrix with unity in the diagonal as a special case. The method, however, can be applied to the unstandardized table of variances and covariances (or square-sums and product-sums) ; and Kelley and others have supposed that the components so obtained would be the same as when obtained from the standardized table of correlations : but it is easy to show that this is not the case. For certain purposes an analysis of the unstandardized covariance matrix seems preferable.

It is instructive to note that the same solution is reached with formulations of the problem that at first sight seem quite different. For example, we may seek (i) weights that minimize the squared discrepancies between the observed *test-measurements* for the individuals and the fit based on the hypothetical factor-measurements, (ii) weights that minimize the squared discrepancies between the observed *correlations* and the fit based on the hypothetical correlations, (iii) weights that maximize the variations between the individuals, i.e., the general *factor variance*, or (iv) weights that maximize the square-sum of the *multiple correlations* between the observed factor-measurements and the hypothetical factor-measurements: all lead to Pearson's method of 'principal components.'

Kendall has stated that "this approach is the first which the statistician would consider, and he [the statistician] may be a little surprised to find that the psychologists do not give it pride of place." But in the early days of factorial work—say, from 1909 to 1924—all those who attempted to factorize complete correlation tables would, I think, have given Pearson's method, if not pride of place, at least priority of place. The reasons for the modifications I have explained elsewhere (16).

The chief objection to the method was, of course, the difficulty of carrying out the calculations; and this, as we have seen, led to the substitution of simpler schemes of weighting and in particular to the use of equal weighting. But, if we may judge from recent discussions of factor analysis by statisticians, the change which has aroused the most widespread criticism is the use of reduced self-correlations. My reasons for making the reduction[1] have been given elsewhere, and may be briefly summarized as follows. With self-correlations of unity we are led (i) to explain the observed correlations by more factors than the data warrant, and (ii) to represent the specific factors by obviously artificial bipolars. (iii) Further, Pearson's procedure, which seeks the 'line of closest fit' by the method of least squares, places all the test-results on the same footing. Now in psychological work the measurements supplied by tests of different kinds are apt to differ widely in their precision and their validity. For example, among the twelve intelligence tests I employed in my first investigation (3), the best had correlations with the criterion amounting to 0·80 or thereabouts; the poorest had correlations of only 0·20 or less. But it is surely unreasonable to insist that the calculated estimates of intelligence should fit the measurements furnished by the poorest tests as closely as they fit the measurements furnished by the best. An allowance for such differences can be made by what I termed 'correction for specificity'; and this in turn entails substituting reduced self-correlations for self-correlations of unity: (see 16, p. 111 and refs.).

III. WEIGHTED SUMS AS FACTORS

A General Method of Factorization. So far our discussion has been restricted to the type of weighting problem most commonly encountered in actual practice, namely, that in which we assume a single dominant factor common to all our tests or question-papers, and are concerned solely to determine a final set of marks measuring this common factor as accurately as possible. Factor analysis, however, goes further than this: it reveals the existence of other factors, assesses their relative importance, and claims to supply a method of estimating measurements for these additional factors, whenever they are required. For the supplementary factors the method of simply summing the scores or their residuals may still be used. But, as can readily be demonstrated, it is likely to be far more inaccurate than for the first factor ; and for later factors the loss of accuracy tends to be cumulative.[1] For rough pilot enquiries, no doubt, the method may often be suggestive.

---

[1] From about 1915 onwards this substitution became almost automatic with investigators who calculated saturations from complete tables of correlations. The procedure more recently proposed by Thurstone may at first sight appear the same as my own: indeed, his brief reference to my views might seem to imply that my 'reduced self-correlations' are identical with his 'communalities' (*Multiple Factor Analysis*, p. ix). But whereas my reduction is determined primarily by considerations of *statistical significance*, Thurstone's is determined by considerations of *rank*. The 'correct values' of the 'communalities,' he tells us, "maintain the rank of the side correlations, and this is the minimum possible rank. . . . If smaller values were taken, the rank would rise" (p. 283). As a result, his values are almost always larger than mine. This will be obvious when we recall that in practice he takes the largest of the observed test-correlations as a convenient approximation to its communality. I cannot see that "the rank of the side correlations" as such can be of any objective importance: with an empirical table it is simply a function of the number of tests we happen to have put into the battery. It is true that the significance of the residuals will in part depend on the number of the tests, but not in the way Thurstone's procedure assumes.

In such cases, however, the question is naturally raised : are we really justified in regarding these rough and ready estimates as representing ' factors ' ? Spearman and those who worked with him insisted that they could not be treated as factors at all ; and this appears to be the general view. But I believe it will not be difficult to prove that such estimations have all the fundamental properties of a factor.

It would, I fancy, now be almost universally admitted that what is commonly called a ' factor ' is essentially a ' weighted sum ' ; yet it is not so widely recognized that any ' weighted sum ' constitutes a ' factor.' To demonstrate this point, however, it will be necessary to define our terms a little more precisely. As Professor Kendall has pointed out, factor analysis has been " developed mainly by men who were not primarily statisticians " ;[2] and in consequence it has " tended to acquire a language of its own." But, although there may be some doubt over minor points, most of the terms employed by psychological factorists can be translated without loss of meaning into the terminology of matrix algebra.

If, then, we agree to regard a table of measurements for n traits obtained from N persons as an $n \times N$ matrix, M_0 say (where $n < N$), we can regard the figures in each row of M_0 as linear functions of r linearly independent $1 \times N$ row-matrices or ' vectors,' chosen to serve as a convenient ' basis ' for all such measurements, where r will in general stand for the rank of M_0 ($r \leqslant n$). Moreover, these selected vectors may or may not be statistically independent. Thus we may write

$$M_0 = FP\,, \qquad \text{(viii)}$$

where F and P denote what (in the usual psychological context) would be called the ' factor-saturations ' for the traits and the ' factor-measurements ' for the persons respectively. The equation states that the measurement for the jth person in the ith trait can be expressed as a weighted sum, viz.,

$$m_{ij} = f_{i1}p_{1j} + f_{i2}p_{2j} + \ldots + f_{ir}p_{rj} = \sum f_{ik}p_{kj}\,, \qquad \text{(ix)}$$

where f_{ik} are the ' weights ' assigned to the kth factor when reconstructing the ith test. By solving the equations for the p's (or adopting some equivalent procedure) we may also succeed in expressing the p's as linear functions or weighted sums of the m's.

Now if we use f_k (with a single subscript) to designate the kth column in F, and p_k to denote the kth row in P, then we can write $f_k p_k = M_k$, say, where M_k will be an $n \times N$ rectangular matrix of ' rank one,' that is, all its rows will be proportional ; in other words, M_k will form what in factorial jargon is called an asymmetrical ' hierarchy.' It follows from (viii) that M_0 can be analysed into r additive sub-tables of this kind, so that

$$M_0 = M_1 + M_2 + \ldots + M_r = \sum M_k\,. \qquad \text{(x)}$$

If equation (x) is to be fulfilled exactly, the number (r) of unit-rank sub-tables in this series will be identical with the ' rank ' of M_0. By the rank of a matrix is meant the number of linearly independent row-vectors (or column-vectors [3]) contained in that matrix ; and, if an exact factorial reconstruction is desired, this will be identical with the total number of 'factors' precisely obtainable from that matrix.

---

[1] I endeavoured to illustrate this conclusion by comparing the results of factor analysis with those of analysis of variance, since with analysis of variance weights are in effect treated as numerically equal (cf. 15, section on ' analysis by simple averaging,' and Table XII, last two lines). The same point emerges if we turn to the illustrative example provided by Professor Kendall. On using factor analysis to extract supplementary factors, we find that these clearly classify his crops into four groups—cereal, leguminous, roots, and grass. But if we use unweighted figures and try to find a possible classification by working with simple ranks and rank-differences, the grouping is greatly obscured.

[2] Cf. 17, § 1. I would also add " and by writers compelled to use language that would be broadly intelligible to psychologists, teachers, and research-workers, who were not themselves familiar with statistical terminology." Thus a non-mathematical psychologist seems to get some inkling of what factorists mean when they use a phrase like ' highest common factor,' whereas he would feel bewildered if they talked of ' the orthogonal component which makes the maximum contribution to the variance ' ; he understands how to calculate a ' weighted sum,' but, were he an editor, he would in the early days have refused (as Ward once did) to publish a paper which spoke of " such technicalities as linear functions and principal axes."

[3] The number will be the same whether we select rows or columns. From this definition it follows that the rank will give the order of the largest minor determinant in M_0 which is not zero. This latter theorem is usually taken as the definition of rank (Bôcher, M., *Higher Algebra*, 1907, p. 22 ; Aitken, A. C., *Determinants and Matrices*, 1939, p. 60). But for purposes of factor analysis the definition given in the text is more convenient : it is that adopted by Macduffee, *Vectors and Matrices*, 1943, pp. 22, 64.

Linearly Independent Factors. The most general way of effecting such an analysis is to take for the first hierarchical matrix an expression of the type

$$M_1 = \frac{1}{v} s_t s_p \, , \tag{xi}$$

where $s_p = w_p' M_0$, $s_t = M_0 w_t$, and $v = w_p' M_0 w_t$. Here s_p will be a row-vector whose elements are derived by weighting and adding the figures in each column of M_0 ; s_t will be a column-vector whose elements are derived by weighting and adding the figures in each row of M_0 (in the usual context s_p will be a row of sums for persons, and s_t a column of sums for tests) ; v, a scalar (analogous to a variance), will be simply an adjustable divisor introduced to bring the products $s_t s_p$ to terms of the same scale as M_0. So far, it is to be noted, w_p' may be *any row of weights*, and w_t may be *any column of weights*, provided, of course, that the divisor v is not thereby reduced to zero.

Now, an application of the usual criterion for rank will quickly show[1] that, whatever be the rank of the initial matrix M_0, the matrix of first residuals, that is, $M_0 - M_1 = M_{0 \cdot 1}$, will *necessarily have a rank which is one degree lower*. Moreover, these residuals in their turn may be factorized in precisely the same way. Once again, any set of weights may be used, provided, of course, they are not the same as the first set ; and once again the rank will be reduced. We can repeat the subtraction and factorization r times (r being the rank of M_0), whereupon the last set of residuals will be found to vanish identically.

At the same time we may note that, at every stage, the standardizing divisor may be considered as applied either to the expression $w_p' M_0$ or to the expression $M_0 w_t$ (and their subsequent analogues) or shared between the two (by taking the square root of the divisor). Thus, according to choice, s_p may be regarded as specifying factor-measurements for the persons (standardized if desired), and s_t the weights for the tests or traits ; or alternatively s_t can be regarded as specifying factor-measurements for the traits (standardized if desired), and s_p the weights for the persons. Consequently, the resolution I have described may be effected with a view *either* to analysing and classifying the persons *or* to analysing and classifying the traits. But, whichever happens to be our aim at the moment, the point I wish here to stress is that *any* set of weights—for instance, weights that are equal to one another (say 1, 1, . . . 1), or a weighting based on a single test (say 1, 0, . . . 0), or even a set of weights chosen absolutely at random—can be used to effect an analysis into factors.

Statistically Independent Factors. A more narrowly defined set of factors can be secured if we select our weights so that the factor-measurements will be not merely linearly, but statistically, independent (i.e., uncorrelated). This can be effected by the simple device of taking $w_t = s_p' = M_0' w_p$, and thus avoiding the introduction of a second set of freely chosen weights. We then have $v = w_p' M_0 s_p = w_p' M_0 M_0' w_p$. Equation (xi) now becomes

$$M_1 = \frac{1}{v} s_t s_p = \frac{M_0 M_0' w_p \, . \, w_p' M_0}{w_p' M_0 M_0' w_p} = \frac{R_0 w_p \, . \, w_p' M_0}{w_p' R_0 w_p} \, , \tag{xii}$$

where R_0 denotes the matrix of observed correlations and M_0 (as usual) is assumed to be in unitary standard measure.

To standardize s_p we must divide it by the root-square-sum. We then have

$$p_1 = v^{-\frac{1}{2}} s_p = \frac{w_p' M_0}{\sqrt{w_p' R_0 w_p}} = w_1' M_0 \text{ (say) ;} \tag{xiii}$$

and we can then take

$$f_1 = v^{-\frac{1}{2}} s_t = \frac{M_0 M_0' w_p}{\sqrt{w_p' R_0 w_p}} = M_0 p_1' : \tag{xiv}$$

so that f_1 is the column of correlations between the test-measurements (M_0) and the factor-measurements (p_1'). We now have for the first hierarchical sub-table $M_1 = f_1 p_1$. As a corollary we have $f_1 = M_0 M_0' w_1 = R_0 w_1$; and therefore $w_1 = R_0^{-1} f_1$. An incidental advantage of this procedure is that it enables us to calculate the factor-saturations f_1 and the factor-measurements p_1 , direct from the small $n \times n$ matrix of correlations, R_0, instead of from the larger initial $n \times N$ matrix of test-measurements M_0.

We can factorize the residual matrix $M_{0 \cdot 1} = M_0 - M_1$ in precisely the same way. We then have for the second set of factor-measurements $p_2 = w_2' M_{0 \cdot 1}$, where w_2' may again be any set of weights. And for the second set of factor-saturations we have $f_2 = R_{0\cdot 1} w_2$. The correlation between the two factor-measurements will now be

---

[1] This statement and the general method of reduction is ultimately due to Lagrange (cf. Bôcher, *loc. cit. sup.*, pp. 131f., and Wedderburn, J. H. M., *Lectures on Matrices*, 1934, pp. 68f.).

$$p_2 p_1' = w_2' (M_0 - f_1 p_1) p_1' = w_2' (M_0 p_1' - f_1) = w_2' (f_1 - f_1) = 0. \qquad \text{(xv)}$$

This result shows that, with *any* two sets of weights, w_1 and w_2, we shall obtain two sets of uncorrelated factor-measurements. (The only restriction is that the divisors $w_1' R_0 w_1$ and $w_2' R_{0\cdot 1} w_2$ shall not be zero ; and this evidently prevents us from taking the *same* set of weights for w_2 as for w_1.)

We can continue the process to analyse the successive sets of residuals, until, after deducting the rth hierarchical sub-table, we shall be left with residuals of zero. It follows that we may (if we wish) choose completely random weights for $w_1, w_2, \ldots w_r$, or (if we prefer) unit-weights such as $[\pm 1, \pm 1, \ldots \pm 1]$ (in which case we have to alter the sign-pattern to prevent $w_1 = w_2 = \ldots = w_r$), or again $[1, 0, 0, \ldots 0]$, $[0, 1, 0, \ldots 0]$, $[0, 0, 1, \ldots 0]$, and so on. We can in this way determine the rank of M_0, and resolve M_0 into an additive series of r hierarchical sub-tables.[1]

In practice, of course, with an empirical set of $n \times N$ test-measurements, the rank of M_0 will be n or N, whichever is smaller ; and our interest will be, not to ascertain the complete number of factors, but the *minimum* number of significant factors, that is, what may be called the 'significant rank' of M_0. Into the further limitations imposed on w_1, w_2, etc., by this or other supplementary conditions it is unnecessary to enter here, since the more specialized procedures then required are already sufficiently well known.

IV. EXPERIMENTAL STUDIES OF DIFFERENTIAL WEIGHTING

An Attempt at Empirical Verification. To illustrate the effects of the different types of weighting we have discussed, it may be helpful to glance at a few concrete cases. The object of the following experiments was to examine instances where there happened to be an objective check on the calculated 'factor-measurements' (the weighted sums). In this way it is not difficult to show that, at any rate so far as the dominant factor is concerned, (i) equal weighting will commonly provide moderately good estimates, (ii) much better estimates can be obtained by a factor analysis, and that (iii) weighted summation with reduced self-correlations will yield better estimates than any other procedure—better even than the classical method of calculating the line of closest fit (i.e., weighted summation with unity in the diagonal).

An Example with an External Criterion. Five children, aged 12 to 15, were given a set of 80 boxes of the same size and shape, but different weight. The weights varied from 150 to 250 grains (about 10 to 16 grams). Each child was asked to assess the weight of the boxes ; and the investigator's problem was to ascertain how closely the estimates derived by averaging or otherwise combining the children's judgments would approximate to the true values, when the mode of combination was determined without reference to those values.

TABLE I. INTER-CORRELATIONS BETWEEN JUDGMENTS OF WEIGHTS

Persons	L. T. R.	F. B.	A. M.	J. C. H.	W. R.
L. T. R. ..	(·931)	·636	·531	·407	·238
F. B.	·636	(·948)	·575	·354	·152
A. M. ..	·531	·575	(·872)	·178	·131
J. C. H. ..	·407	·354	·178	(·706)	·053
W. R. ..	·238	·152	·131	·053	(·553)

---

[1] The last two types of unit weighting are of special interest. The first is the method of weighting implicitly adopted in analysis of variance : (see Burt, 16, p. 7). The second is a simplified form of the old triangular reduction discussed by Lagrange and Hermite. It is particularly useful as a preliminary reduction to simplify subsequent calculations. It also has a practical application in cases where (after suitable rearrangement) later variables are dependent on earlier but earlier are not dependent on later : (see Burt, 10, p. 307 and refs.). Its introduction into correlation work is due indirectly to Karl Pearson (cf. *Biometrika*, XV, p. 135, and XXIV, p. 422).

As with Galton's test-weights, the successive increments were geometrical[1]; and the set was so selected that, judged by their logarithmic values, the frequency-distribution of the weights would approximate to a normal curve. They were presented in a randomized order. Before lifting the first, each subject was given three standard boxes which he was told weighed 100, 200, and 300 grains. To allow for Weber's law both true values and assessments were converted to terms of a logarithmic scale, and were then reduced to standard measure.[2]

Owing partly to the comparatively small differences in the weights and partly to differences in care or discriminative capacity among the individual children, the agreement between their judgments is by no means close. This is shown by the inter-correlations,[3] which are set out in Table I, and give what may be called the 'inter-personal reliability' of the assessments. With 80 items the standard error of a zero correlation is approximately ± 0·112, and any coefficient over 0·217 may be accepted as significant. To obtain a measure of personal error or 'intra-personal reliability' for each child, the experiment was repeated with each subject after an interval of a week; and a correlation calculated between his judgments as given at two successive trials—a procedure originally suggested by Clark Wissler.[4] To check the weighting equations derived from the first experiment the same children repeated the test with a second set of boxes ranging from 200 to 300 grains.

Agreement of Individual Assessments with the Criterion. (a) *External Criterion.* Each child's assessments were first correlated with the 'true weights.' The coefficients are shown in the first column of Table IV. It will be seen that the correlations range from 0·196 (which is scarcely significant) to 0·841. Two out of the five children not only have reliabilities well over 0·90, but also appear able to arrange the boxes very satisfactorily in order of weight. Indeed, quite a good estimate would be obtained by accepting the judgments given by L. T. R. alone.

It would, however, seem reasonable to anticipate a more accurate estimation, if we relied, not on the judgments of a single individual, however good, but on the composite 'index' (as Pearson would have called it), which can be obtained by averaging the judgments of two or more individuals. Indeed, it is commonly supposed that the larger the number of assessments available for averaging, the greater is the accuracy of the average so obtained. First of all, therefore, let us try averaging the judgments obtained from all the judges. On calculating a straight average for each of the 80 boxes, we obtain a correlation with the true values of only 0·844. (The same coefficient is reached if we use equation (vi) above.) This figure exceeds that obtained with the best of the five judges by an insignificant fraction only; and is not nearly as high as that which would be obtained by averaging the assessments of the best two judges (0·902). It is clear that, by averaging all the assessments indiscriminately, we have allowed the inaccuracies of the poorer subjects to counterbalance the gain to be expected by pooling the judgments of the best. However, in investigations where we did not know the true values, we could not at this stage say who really were the best judges.

How, then, in the absence of any external check, can we determine their accuracy? Apparently our only means of doing so is to take the joint report of all the judges as a provisional indication, and then measure the accuracy of each one by his agreement with this internally derived criterion. We could then improve our provisional indication by substituting a weighted average on the basis of this result; and we could continue by successive approximation until we reached a stable set of figures. Factor analysis is merely a short cut to this final result.

---

[1] Cf. Galton, F., *Inquiries into Human Faculty*, 1883, Appendix, pp. 248–51.

[2] The experiment was suggested by the test of 'five weights' introduced into the original Binet-Simon Scale (cf. Burt, C., *Mental and Scholastic Tests*, 1921, p. 51, and 'Correlations between Persons,' *Brit. J. Psych.*, XXVIII, 1937, pp. 63-4). In practical laboratory courses for students, both at Liverpool and London, I have regularly used it as a stock exercise (usually with 10 weights only) to illustrate various correlational methods, and, in particular, the analogies and the differences in the procedure required (*a*) when we possess an external criterion and (*b*) when we have to rely only on an internal criterion. Incidentally it exemplifies the application of factor analysis to correlations between persons as well as between tests or traits.

[3] The original calculations were carried out by Mr. R. F. Chalmers, whom I have to thank for permission to use his figures; and most of the work has been checked by Mr. Summerfield, to whom I am also much indebted. I should add that more than five children took the test; but the correlation table was limited to that number to simplify both the exposition and the work.

[4] 'The Correlation of Mental and Physical Tests,' *Psych. Rev. Mon. Supp.*, III, 1901, p. 60. In all our earlier experiments with mental tests it was part of the regular routine to repeat the tests, and calculate the reliability coefficients. In theory a more satisfactory derivation of the formulæ for factor analysis becomes possible, when values for the errors are assumed to be known; but in practice it is doubtful whether the estimates would thereby be appreciably improved, particularly when only one factor has to be determined (cf. 3, pp. 112, 157).

(b) *Internal Criterion.* Let us now imagine that we present a list of figures for the five sets of standardized assessments to an investigator who knows nothing whatever about their actual nature or source (he might, for example, suppose they were marks for 80 pupils in five subjects of the elementary curriculum); and let us ask him to determine measurements for as many independent factors as will reasonably account for the variation in the figures.

It is evident that he could not in any case determine measurements for *more* than five common factors. If he follows Pearson's procedure, he will take the variances or self-correlations to be unity, calculate his factor-saturations by weighted summation, and so obtain weighting equations for the maximum number of components, namely, five: the matrix of regressions or weights for determining the measurements required will then be simply the inverse of the factor-matrix. The factor-saturations thus obtained are shown in Table II A. It will be seen that the first component accounts for about 49 per cent. of the total variance, and that the other components account for 19, 17, 8, and 7 per cent. respectively.

TABLE II. SATURATION COEFFICIENTS

Persons	A. For Observed Table					B. For Smoothed Table				
	I	II	III	IV	V	I	II	III	IV	V
L. T. R. ..	·861	·005	·018	−·307	−·410	·865	−·062	−·138	−·268	−·398
F. B.	·849	−·096	−·156	−·255	·428	·835	−·070	−·156	−·333	·402
A. M.	·754	−·020	−·481	·440	−·066	·747	−·097	−·304	·585	·032
J. C. H. ..	·556	−·354	·717	·221	·036	·550	−·201	·803	·106	·011
W. R.	·321	·907	·252	·073	·062	·316	·940	·118	·040	·007
Variance ..	2·443	·957	·834	·406	·360	2·405	·942	·794	·538	·321

TABLE III. CHI-SQUARED ANALYSIS[1]

(i) k	(ii) Latent Root (λ)	(iii) R_q	(iv) $\text{Log}_e R$	(v) χ^2 ($74{\cdot}5 \log_e R_q$)	(vi) D.F.	(vii) Border-line	(viii) χ^2 (Diffs.)	(ix) D.F.	(x) Border-line
0	2·443	·2870	−1·2483	93·0	10	18·3	66·8	4	9·5
1	·957	·7037	−0·3514	26·2	6	12·6	10·2	3	7·8
2	·834	·8072	−0·2142	16·0	3	7·8	15·8	2	6·0
3	·406	·9974	−0·0026	0·2	1	3·8	0·2	1	3·8
4	·360	1·0000	−0·0000	0·0	0	—	—	—	—
Total	5·000	—	—	—	—	—	93·0	10	—

---

[1] *Note to Table III.* In col. (i), k denotes the number of factors that have been extracted. Write $q = p - k$ for the number of roots still to be extracted. Then in col. (iii),

$$R_q = q^q \frac{\lambda_{k+1}\,\lambda_{k+2}\,.\,.\,.\,\lambda_p}{(\lambda_{k+1} + \lambda_{k+2} + \,.\,.\,.\, + \lambda_p)q} = q\text{th power of } \frac{\text{Geom. mean of last } q \text{ roots}}{\text{Arith. mean of last } q \text{ roots}};$$

col. (v) gives chi squared for the residual correlations remaining after k factors have been removed; the multiplier $= N - p - \frac{1}{2} = 80 - 5 - \frac{1}{2} = 74{\cdot}5$, where $N =$ the number of items over which the correlation extends, and p the number of persons correlated; col. (vi) gives the corresponding degrees of freedom, and col. (vii) the borderline values of chi squared at the 5 per cent. level; col. (viii) gives chi squared for the separate factors (obtained by taking successive differences from col. (v)); and cols. (ix) and (x) give the corresponding degrees of freedom and borderline values.

I am indebted to Professor Bartlett for an advance note giving his formula for testing significance. His method has been described more fully in his paper published in this issue (see pp. 77–85). My own method of expressing his formula was intended partly to suggest a routine method of computation when all the latent roots are known, and partly to indicate certain analogies with other statistical tests (e.g., for homogeneity of variance). Note that, when the side correlations are approximately hierarchical, $|R|$ can be calculated by the formula $|R| = (1 + f'U^2f) \cdot |U^2|$, where $f =$ the vector of saturations, and U^2 the diagonal matrix of specific factor variances.

Now with a ' principal axes ' analysis, based on the total variance (unity in the diagonal of the correlation), we can apply the chi-squared test in the form suggested by Bartlett (Table III). From the figures it will be seen that three factors at most would be sufficient to account for all that is significant in the correlation structure. Further, on examining the pattern[1] presented by the factor-saturations, it becomes clear that the four bipolar factors can represent little else than the results of forcing specific factors into the form of common factors. This can best be shown if we reconstruct a smoothed correlation matrix from the saturations for the first factor alone, and then factorize it by Pearson's method with unity in the diagonal : the results are given for comparison in Table II B, where the artificial nature of the resulting factor pattern is very plain. In Table II A there is certainly a suggestion of a small but genuine bipolar factor contrasting F. B. and A. M. with the other three subjects ; but, until this has been disentangled from the effects of the specific factors, its significance must remain highly doubtful. Our investigator can therefore, with little or no hesitation, assume the existence of a dominant factor with positive saturations throughout, but that is all ; and the hypothesis which at this stage appears to be at once the most economical and the most plausible will be that the five sets of correlated assessments are explicable as the result of (i) a single general factor and (ii) five specifics.[2] Only if this factor-structure is found unavoidably to leave residuals that are definitely significant shall we be warranted in postulating (iii) one or more supplementary factors in addition.

His next step, therefore, will be to test his hypothesis by finding how closely he can fit his data with saturations deduced from it. To find the saturations yielding the closest fit he will again use the method of weighted summation, but will now insert appropriately reduced self-correlations into the diagonal of the correlation matrix.

The principle of economy demands that he should begin by basing his estimated self-correlations on the assumption of a single common factor first of all, and only introduce further factors when the evidence obliges him to ' multiply the entities ' he is postulating. Table I can be fitted reasonably well with one common factor alone. The largest residual is $-0{\cdot}088$; and this can have no statistical significance. Had the first set of residuals been statistically significant (as it might easily have been with a larger group), then he would have to augment his self-correlations accordingly, and commence his calculations afresh. In cases like the present, where probable errors are fairly large, he might in actual practice be content with simple summation instead of weighted. On the other hand, if he wished to adopt a more rigorous treatment, he would first ' correct for unreliability ' (thus taking a matrix of maximum reliability) ; or, if the reliabilities were unknown (as is more usually the case), ' correct for specificity ' (16, p. 113 ; the formulæ are the same, except that in the former case r_{ii} denotes reliabilities, in the latter reduced self-correlations).

Our investigator's ultimate aim, however, is to estimate the appropriate measurements for the first (and only significant) factor. This factor provides us with a criterion based on internal evidence ; and, so far as it can be accepted as a trustworthy substitute for the external evidence of the balance, we may take the ' saturation coefficients ' for this factor as estimates of the probable correlation of each judge's assessments with the ' true weights.'

Accordingly, we have next to consider the problem of finding the most suitable weights for this purpose. Let us compare the results furnished by the various methods available for calculating (*a*) these estimated correlations between the assessments and the criterion, and (*b*) the regression coefficients to which these correlations would severally lead.

Judges' Correlations with the Criterion. The saturation coefficients are shown in the last five columns of Table IV.

---

[1] Cf. 16, p. 112, for a fuller explanation of this point. I have to acknowledge the assistance of Mr. Summerfield, who has been good enough to check my calculations.

[2] Had he been allowed some antecedent knowledge of the source, this is the factor-pattern that I imagine he would regard as most probable on *a priori* grounds. If the data had been procured as a result of some hypothesis that he himself had set up (the ideal procedure in a factorial study), that hypothesis, I imagine, would have begun by postulating a three-factor scheme—with basic, group, and specific factors ; his task would then have been to verify or disprove this hypothesis either wholly or in part, and, if it were wholly verified, to measure the relative importance of the several factors by calculating the most likely estimates for their variances. The bipolar or group factors would, of course, represent the contrast between different types of judge—e.g., those who are more or less subject to weight-illusions, or possibly Binet's *interprétateurs* and *simplistes ;* and the judges might even have been deliberately chosen in the hope of eliciting such type-differences if they existed (for the existence of such factors, see *Brit. J. Educ. Psych.*, XIX, 1949, p. 110, and refs.).

Five methods of analysis have been employed. (i) Weighted summation, with self-correlations of unity (*W.S.u.s.*); this, it will be remembered, adopts as the internal criterion Pearson's line of closest fit. (ii) Simple summation, with self-correlations of unity (*S.S.u.s.*); here the criterion is the straight unweighted average of the standardized assessments. (iii) Weighted summation with correction for specificity (*W.S.c.s.*). (iv) Weighted summation with reduced self-correlations, but without correction for specificity (*W.S.r.s.*). (v) Simple summation with reduced self-correlations (*S.S.r.s.*).

TABLE IV. AGREEMENT OF INDIVIDUAL ASSESSMENTS WITH CRITERION

Persons	(*a*) With External Criterion	(*b*) With Internal Criterion (Saturation Coefficients)				
		(i) *W.S.u.s.*	(ii) *S.S.u.s.*	(iii) *W.S.c.s.*	(iv) *W.S.r.s.*	(v) *S.S.r.s.*
L. T. R. ..	·841	·861	·829	·818	·838	·863
F. B.	·790	·851	·801	·805	·810	·815
A. M. ..	·634	·753	·712	·669	·642	·625
J. C. H. ..	·402	·556	·587	·427	·414	·399
W. R. ..	·196	·320	·464	·238	·221	·214
Discrepancy (Root-mean-square)	·000	·107	·149	·030	·015	·018

TABLE V. WEIGHTING COEFFICIENTS (PARTIAL REGRESSIONS)

Persons	(*a*) Based on External Criterion	(*b*) Based on Internal Criterion				
		(i) *W.S.u.s.*	(ii) *S.S.u.s.*	(iii) *W.S.c.s.*	(iv) *W.S.r.s.*	(v) *S.S.r.s.*
L. T. R. ..	·534	·353	·295	·421	·480	·532
F. B.	·346	·350	·295	·378	·400	·403
A. M. ..	·154	·308	·295	·204	·141	·101
J. C. H. ..	·040	·228	·295	·081	·048	·022
W. R. ..	−·002	·130	·295	·048	·029	·011
Correlation						
(i) Main Set	·914	·885	·844	·907	·910	·908
(ii) Second Set	·862	·781	·837	·885	·902	·873

Let us first enquire how far the hypothetical correlations with the internal criterion afford a trustworthy indication of the judges' observed correlations with the external criterion, i.e., the true weights. The closeness of the approximation can best be assessed by computing the root-mean-square of the several discrepancies (see last line of Table IV). It is plain that the nearest figures are the saturations furnished by the method of weighted summation with reduced self-correlations. Those found by taking self-correlations of unity are decidedly inaccurate: the effect of this procedure is to give a deceptively high figure for the accuracy of the poorest judges (especially J. C. H. and W. R.). The worst figures of all are those furnished by simple summation with self-correlations of unity.

Estimation of the Criterion by a Weighted Combination of the Judges' Assessments. As we have seen, neither the assessments of the most accurate judge nor yet a simple average of all the assessments can be expected to yield the best available estimates for the weights of the several boxes. Our problem now is to enquire how far such estimates are improved by the different methods of weighting that are available. For this purpose the necessary weights or loadings will be calculated from the correlations or saturations by the usual formula for partial regressions.

If we already knew the true values for each box, we could calculate the regressions from the observed correlations: the equation to use would be $w' = r'R^{-1}$, where r stands for the first column in Table IV and R the correlation matrix in Table I with unity in the diagonal. The loadings so obtained are set out in the first column of Table V. The figures indicate that A. M.'s judgments deserve less than half the weight of F. B.'s and little more than a quarter of that of L. T. R.'s, while the assessments of J. C. H. and W. R. have no significance at all. In general, of course, the regressions obtained in this way from a direct comparison with the true values will furnish the highest multiple correlation. Here, it will be seen, the multiple correlation amounts to 0·918.

Now, as I ventured to suggest in my earlier papers (4, cf. 10 and refs.), even where we have no external criterion we can still apply the principles of multiple correlation to determine measurements for our so-called 'factors.' The regressions will be calculated by a similar formula ($w' = f'R^{-1}$, where f' will now denote any of the last five columns of Table IV). And generally, to determine the relations between the internal criteria (P) and an external criterion (c), we may use the same weights as those I have elsewhere suggested for computing the correlations between a set of factors and *any* external variable: viz., (in the notation of my previous papers) $W = F^{-1}$ (if F is non-singular), or $F'R^{-1}$ (if it is not), i.e., $V^{-1}F'$ (if weighted summation has been used), or $v^{-1}f'$ (if only one factor is required). To transform F to r_{mc}, we adopt the usual type of post-multiplier, viz., $F^{-1}r_{mc} = F^{-1}Mc' = Pc' = r_{pc}$, making the same substitutions as before if F is non-singular (cf. 10 and Appendix, p. 124). The weighting coefficients thus obtained are shown in the corresponding columns of Table V. The correlations with the *true* values, calculated by using these weights, are appended in the last line but one. Confirmatory validation, as usual, has been sought by trying the same weighting equations with a *second* sample: see last line of table.

It is clear that simple averaging (*b* ii) yields the least satisfactory result. As we have seen, it gives equal weights to all judges. Even the Pearsonian method for finding the 'line of closest fit' (*b* i) tends in the same direction. On the other hand, the methods of simple and weighted summation *with reduced self-correlations* (*b* iii, iv, and v) here supply almost as close an agreement with the true values as was obtained from the partial regressions based on the true values themselves. In the present case the differences between the last three methods are too slight for any generalization to be deduced from the figures alone.

Supplementary Experiments. To show that these results are in no way exceptional, but hold good in widely different fields, I append the figures calculated by similar methods from six other enquiries.

The tasks were (1) to judge the weights of boxes (an experiment identical with the above but carried out with adults); (2) to judge the length of lines; (3) to judge the areas of rectangular figures; (4) to judge the length of time-intervals; (5) to judge the number of spots on cards shown tachistoscopically; (6) to guess the heights of familiar buildings or monuments. In the first and last of these experiments adults (students) acted as judges; in the other three, children aged 12 to 14. Table VI shows the final correlations between the 'true values' and the estimates given by various linear combinations of the assessments made by all the judges.[1]

TABLE VI. CORRELATIONS BETWEEN JUDGMENTS AND TRUE VALUES

Test Material	No. of Items	No. of Judges	(*a*) Based on External Criterion	(*b*) Based on Internal Criterion (i) *W.S.u.s.*	 (ii) *S.S.u.s.*	 (iii) *W.S.c.s.*	 (iv) *W.S.r.s.*	 (v) *S.S.r.s.*
Weights ..	30	10	·963	·954	·937	·966	·958	·972
Lines ..	50	6	·927	·813	·748	·913	·930	·895
Areas ..	20	8	·668	·623	·605	·545	·584	·491
Time-intervals	50	6	·751	·680	·663	·738	·742	·731
Spots ..	100	7	·842	·776	·736	·791	·833	·802
Buildings ..	20	8	·765	·719	·721	·754	·772	·768
Average ..			·819	·761	·735	·785	·803	·777

---

[1] The experiments with children were carried out and analysed by Mr. R. F. Chalmers, Mr. B. C. Murray, and Miss K. Willis, to whom I am indebted for the figures here given. The number of judges is in every case rather small: that is almost inevitable, owing to the labour involved in calculating an inverse for larger correlation tables.

The results may be summarized as follows. Where the number of items is small, or where the number of judges is comparatively large, and particularly where all the judges have much the same degree of efficiency, there is little to choose between the various methods of weighting, and not much is lost if no weighting whatever is employed. But where the number of items is large enough to reduce the probable error to small dimensions, and particularly where the number of assessments to be averaged is small, there the use of an accurate method of weighting is plainly of special importance. In such cases the best procedure would seem to be the method of least squares (weighted summation) with reduced self-correlations. The method of simple summation usually yields very similar results. There is one exception—the experiment with areas. Here (as the introspections show) at least one large supplementary factor is almost certainly operative (representing subjective differences in mode of assessment). With so few cases the residuals for such a factor were not statistically significant, and tended to be somewhat erratic. However, if a second factor were allowed for in the self-correlations, these would approach nearer to unity. Hence the apparent exception is in keeping with the general rule. It may seem surprising that the correction for specificity does not (except in one dubious case) improve the resulting figures. That is doubtless due to the fact that the external criterion is not absolutely identical with the internal.

Is Weighting Really Necessary? In discussing the weighting of test items for validity, Guilford apparently concludes that " what few results we have point definitely to the inference that such weighting is not worth the trouble " (*loc. cit.*, p. 448). Can we then lay down any criterion for deciding when, if ever, such calculations are really worth while? I would suggest, as a sound practical rule, that differential weighting may be regarded as unnecessary when the correlation between the weighted and unweighted assessments is significantly larger than the reliability coefficient for the weighted assessments.

Thus, to take the first experiment described above, we find that, with method (*b* iv) (weighted summation) the reliability of the composite figures would be 0·969 ; the correlation between these composite figures and those given by the unweighted average of the standardized judgments is 0·941. In this case, therefore, the use of differential weighting would certainly seem desirable. Whenever the number of tests to be averaged is small, and more especially when their inter-correlations are low, the problem of weighting ought in my view never to be dismissed without at least a rough preliminary investigation of the point.

V. SUMMARY AND CONCLUSIONS

1. An attempt has been made to investigate the conditions affecting the results obtained with different methods of weighting. The problem leads to a consideration of (*a*) *a priori* weighting as the oldest form, (*b*) random weighting as the most general form, (*c*) equal weighting as the simplest form, and (*d*) differential weighting as determined empirically by various factorial procedures.

2. *A priori* or subjective weighting may be necessary (i) where questions of value are concerned, or (ii) where the criterion to be predicted is genuinely composite. Where the criterion is factual and essentially unitary (though assessed by multiple observations), *a priori* weighting is generally worthless.

3. Equations have been derived to express the correlations obtained both by random weighting (such as tends to be produced by accidental differences in standard deviations) and by equalized weighting (where such accidental differences have been eliminated). The formulæ indicate that the validity coefficients so obtained increase (i) with the number of weighted traits and (ii) with the size of the inter-correlations between them. It is concluded that the choice of the best possible set of differential weights is of special importance when the traits are few in number and differ widely in their diagnostic value.

4. A series of illustrative experiments is described in which an external as well as an internal criterion was available. The results indicate that, when an external criterion is not available, the best weights will be obtained with a factor analysis carried out, not by the classical method giving the 'lines of closest fit' (i.e., the method of principal components with self-correlations of unity), but by weighted summation with reduced self-correlations. Where the correlation matrix does not depart

appreciably from a rank of one, the further corrections for specificity or reliability introduce no discernible improvement, and simple summation gives almost as good a result as weighted summation.

5. More generally it has been shown (i) that *any* method of weighting will yield a ' factor ' in the accepted sense of the term, and that the extraction of such a factor always reduces the rank of the matrix of test-measurements and the matrix of the correlations (or the residual matrices) by one ; (ii) that the choice of a weighting system which yields *orthogonal* factors leads to particularly simple formulæ and convenient working procedures. One such system is obtained if the weights selected are numerically equal, but differ in sign-patterns for successive factors, i.e., if the standardized measurements are merely summed *algebraically* as they stand. This is the weighting adopted in analysis of variance. When the number of traits to be summed is small, this type of weighting does not yield altogether satisfactory results ; but when the number is large, it yields a good approximation to the best result, particularly for the first or ' general ' factor.

6. The weighting coefficients deduced from a particular sample necessarily give better predictions for that sample than for any other, especially when the number of predictors is large. Hence an empirical weighting equation should always be checked by application to a second experimental group. When this is done, weighted summation with reduced self-correlations again seems to furnish the most stable results.

7. A criterion has been suggested, based on the reliability of the composite result, to indicate when differential weighting should be substituted for equal weighting.

REFERENCES

1. Edgeworth, F. Y. (1888). ' The statistics of examinations,' *J. Roy. Stat. Soc.*, LI, 599–635.
2. Johnson, W. W. (1892). *The Theory of Errors and the Method of Least Squares*. Oxford : University Press.
3. Burt, C. (1909). ' Experimental tests of general intelligence,' *Brit. J. Psych.*, III, 94–177.
4. Burt, C. (1917). *Distribution and Relations of Educational Abilities*. London : King.
5. Bowley, A. L. (1920). *The Elements of Statistics*. London : King.
6. Burt, C. (1921). *Mental and Scholastic Tests*. London : King.
7. Keynes, J. M. (1921). *A Treatise on Probability*. London : Macmillan.
8. Edgeworth, F. Y. (1925). ' Index numbers.' *ap*. Palgrave, *Dictionary of Political Economy*.
9. Brunt, D. (1931). *The Combination of Observations*. Cambridge : University Press.
10. Burt, C. (1935). ' Memorandum on Methods of Factor Analysis ' ; subsequently published in Hartog, P., and Rhodes, E. C. *The Marks of Examiners*, pp. 245–314. London : Macmillan.
11. Monroe, W. S., and Engelhart, M. D. (1936). *The Scientific Study of Educational Problems*. London : Macmillan.
12. Frazer, R. A., Duncan, W. J., and Collar, A. R. (1938). *Elementary Matrices*. Cambridge : University Press.
13. Kendall, M. G. (1939). ' The geographical distribution of crop productivity in England,' *J. Roy. Stat. Soc.*, CII, 21–48.
14. McClelland, W. (1942). *Selection for Secondary Education*. London : University of London Press.
15. Burt, C. (1948). ' A comparison of factor analysis and analysis of variance,' *Brit. J. Psych., Stat. Sect.*, I, 3–26.
16. Burt, C. (1949). ' Alternative methods of factor analysis, and their relation to Pearson's method of principal axes,' *Brit. J. Psych., Stat. Sect.*, II, 98–121.
17. Kendall, M. G. (1950). ' Factor analysis as a statistical technique,' *J. Roy. Stat. Soc.* (in the press).

APPENDIX

1. In the text (p. 108) I have contended that the real or operative weight of any component is to be measured (as in factor analysis) by its contribution to the total variance. Thus, if

$$y = B_1X_1 + \ldots + B_nX_n = B_1\sigma_1x_{1i} + \ldots + B_n\sigma_2x_{ni} = U_1x_1 + \ldots + U_nx_n\,,$$

where $B_j =$ the *assigned* weight and $U_j = B_j\sigma_j =$ the *actual* weight, then

$$\begin{aligned}\sigma_y{}^2 &= B_1\sigma_1\sum B_i\sigma_ir_{1i} + \ldots + B_n\sigma_n\sum B_i\sigma_ir_{ni}\\ &= U_1\sum U_ir_{1i} + \ldots + U_n\sum U_ir_{ni}:\end{aligned}$$

so that the contribution of j depends on (i) B_j, (ii) σ_j, and (iii) the r_{ij}'s. If the components are in standard measurement and uncorrelated, then $\sigma_y{}^2 = U_1{}^2 + \ldots + U_n{}^2$, i.e., the contribution of each component varies with the square of its weight. If the components are correlated, and given equal weights of unity, then $\sigma_y{}^2 = \sum r_{1j} + \ldots + \sum r_{nj} = r_{1y} + \ldots + r_{ny}$, i.e., the contribution of each component varies with its correlation (or 'saturation') with the composite measurement, y. But with differential weighting the saturations will not be proportional to the regressions; and sometimes, as with 'suppressor' variables, a test with a positive saturation may require a negative weight.

2. Where the final criterion is external to the battery, the best method of weighting is that described in the theory of multiple correlation. Where no such criterion is available, then (as stated above) the same kind of proof leads to the formulæ for factor analysis by weighted summation. But, it may be asked, when *both* are available, how are we to determine the relation between the two?

There are two ways in which the relation between a factor and an external criterion may be assessed. We may correlate either (i) the factor-*measurements* (p) with the criterion *measurements* (c), or (ii) the factor-*saturations* (f') with the correlations between the criterion and the several assessments (r_{mc}). The former is in my view the proper procedure; the latter provides a convenient short cut, but has sometimes been adopted without a real understanding of its implications.[1]

(i) If p is in unitary standard measure, then the correlation between the measurements will be $r_{pc} = pc' = w'Mc' = v^{-1}f'r_{mc} = v^{-1}f'r_{mc} = (f'f)^{-1}f'r_{mc}$. But if p consists of *estimated* measurements, then its variance will be $w'f = f'R^{-1}f$; and this is only unity when f has been derived by factorizing the correlation table with self-correlations of unity. Hence the formula used in the text is $w'r_{mc} \div \sqrt{w'f}$.

(ii) The 'unadjusted correlation' between the saturations and the criterion correlations (r_{mc}) will be $\dfrac{f'r_{mc}}{\sqrt{(f'f)}\sqrt{(r_{mc}'r_{mc})}} = v^{-1}f'r_{mc}$, if we take $r_{mc}'r_{mc}$ as $= f'f$ approx. It will be seen that, with this assumption, the two methods of calculation lead to the same result. Moreover, when the 'unadjusted correlation' $= 1$, then the saturations and criterion correlations are proportional. If, however, we desire to compare *several* sets of saturations (e.g., sets obtained by different procedures), then it would seem better to use the square-sum of the differences, though this is again closely related to the unadjusted correlation.

3. The full proof of the equations (iii) and (iv) given on p. 111 was given in my original (unpublished) memorandum (cf. 10, p. 246): though lengthy, it is straightforward. The reader will find no difficulty in reconstructing the argument for himself. The essential steps may be summarized as follows.

Using the notation suggested on p. 111 consider the product-moment correlation between the weighted sums, y and z, defined by equations (i) and (ii); write the correlation $r_{yz} = \dfrac{N}{D_u D_w}$, where N and D_uD_w denote the numerator and denominator respectively.

Then, since the measurements are in unitary standard measure,

$$N = \sum_i^n U_iW_i + \sum_{i \neq j}^{n(n-1)} U_iW_jr_{ij}\,.$$

Putting $U_i = \overline{U} + u_i$ and $W_i = \overline{W} + w_i$, we have

$$\sum_i^n U_i\,W_i = n\,\overline{U}\;\overline{W} + \sum u_iw_i\,.$$

Now U_i and W_i, having been chosen at random, must have only non-significant correlations with each other and with the values they are used to weight. Further, unless there are wide divergences among the inter-correlations, we may take $\sum U_i\,W_j\,r_{ij} = \bar{r}\sum U_i\,W_j$ approximately (where $\bar{r}$ denotes the average value of r_{ij}, $i \neq j$). Hence $\sum\limits_{i \neq j}^{n(n-1)} U_iW_jr_{ij} = n(n-1)\,\overline{U}\,\overline{W}\bar{r} + \bar{r}\sum u_i\,w_u$ approximately.

---

[1] For the formulæ for calculating the correlation between factors and external variables, see 10: also *J. Psychol.*, VI, 1938, p. 355, eq. (vii)–(ix).

But, for the same reason as before, $\sum u_i w_j$ must tend to zero. Hence

$$N = n\,\overline{U}\,\overline{W}\,(1 + n - 1\,\bar{r}) \text{ approximately.}$$

Similarly $D_u = n\,\{\overline{U}^2\,(1 + (n-1)\,\bar{r}) + (1 - \bar{r})\,\sigma_u^2\}$, with a corresponding expression for D_w.

Now, the correlation r_{yz} depends solely on the *relative* proportions of the weights ; hence it will be unaffected if we multiply the first set of weights by some constant, so that $\overline{U} = \overline{W}$. Then, since by hypothesis $\sigma_y^2 = \sigma_z^2 = 1$, we have $\sigma_u^2 = \sigma_w^2$; and therefore

$$\begin{aligned} r_{yz} &= \frac{n\overline{W}^2\,(1 + n - 1\,\bar{r})}{n\,\{\overline{W}^2\,(1 + n - 1\,\bar{r}) + (1 - \bar{r})\,\sigma_w^2\}} \\ &= \frac{n\,\bar{r} + (1 - \bar{r})}{n\,\bar{r} + (1 - \bar{r}) + (1 - \bar{r})\,\dfrac{\sigma_w^2}{\overline{W}^2}}\,. \end{aligned}$$

Dividing out we obtain

$$r_{yz} = 1 - \frac{(1 - \bar{r})}{n\,\bar{r}} \cdot \frac{\sigma_w^2}{\overline{W}^2} + \text{terms in } \left(\frac{1}{n}\right)^2 \text{ and higher powers.}$$

If we prefer to retain the original values for the weights, so that $\overline{U} \neq \overline{W}$, and $\sigma_u \neq \sigma_w$, then $D_u D_w$ can be factored by 'completing the square'; and we reach the same general expressions as before except that for $\frac{\sigma_w^2}{w^2}$ we substitute $\frac{1}{2}\left\{\frac{\sigma_u^2}{\overline{U}^2} + \frac{\sigma_w^2}{\overline{W}^2}\right\}$.

If we suppose that the inter-correlations are too widely divergent to substitute $\bar{r}\sum U_i W_j$ for $\sum U_i W_j r_{ij}$, the algebra becomes somewhat heavier ; but it is not difficult to show that the effect is merely to diminish the value of the negative term, so that it tends towards $\frac{1}{n\,\bar{r}} \cdot \frac{1}{2}\left\{\frac{\sigma_u^2}{\overline{U}^2} + \frac{\sigma_w^2}{\overline{W}^2}\right\}$.

4. The other equations given in the text are readily reached by the use of matrix notation. The following results may be noted. If y and z represent factor-measurements, and the vectors of weights, w_y and w_z (say), are based on factor-saturations, f_y and f_z, then

$$r_{yz} = \frac{w_y'\,R\,w_z}{\sqrt{w_y'\,R\,w_y \,.\, w_z'\,R\,w_z}} = \frac{f_y'\,R^{-1}\,f_z}{\sqrt{f_y'\,R^{-1}\,f_y \,.\, f_z'\,R^{-1}\,f_z}}\,.$$

If w_y has been obtained by a Pearsonian analysis, $r_{yz} = f_y'\,w_z$, assuming w_z is standardized.

If, however, $w_z = [1, 1, \ldots 1] = w_0$ (in the usual notation), then $r_{yz} = \dfrac{f_y'\,w_0}{\sqrt{w_0'\,R\,w_0}}$.

NOTE ON SIR CYRIL BURT'S PAPER ON DIFFERENTIAL WEIGHTING

By GODFREY THOMSON
Moray House, University of Edinburgh

I fear the preceding paper may be misleading unless it is read with the following warning in mind. I am thinking more particularly of the experiments with lifted weights, etc., in which it was found that the method of weighted summation *with reduced self-correlations* provides almost as good an agreement with the true weights as would be obtained from the partial regressions based on those true weights ; and the method of simple summation provides estimates that are almost as good (pp. 121, 123). The conclusion implicitly drawn by the writer is that ' reduced self-correlations,' which most of us call ' communalities,' are justifiably used in analyses of psychological tests—a conclusion about which I am very sceptical. In the experiment with lifted weights we *know* that the correlations are entirely due to one factor, the weight, which is the only thing common to the five persons making the judgments. In that case, it seems to me obvious that it is proper to reduce the rank of the matrix, by the insertion of communalities, until it is of rank one, or as nearly rank one as the inevitable blemishes will permit. Burt's experimental confirmation of this is interesting, but the result seems to me to be a foregone conclusion. To reverse this argument, however, seems to me to be dangerous. In a psychological set of correlations between tests (and the persons in Burt's experiments correspond to tests in the usual psychological analysis, while the weights in his experiment correspond to the children in the usual psychological experiment), even when the rank can be reduced by communalities to one, we do not know that there is only one factor causing the correlations. The correlations may have been caused in many other ways, and in particular might have been caused by tests which call upon random samples of a large number of qualities whether inherited or acquired (genes, bonds, pieces of information, fragments of skill, what you will). It is true that when the rank can be reduced to one, the tests have acted " as if " they had only one factor in common, together with a large specific factor in each test, but this is not necessarily the case, and it seems to me that to insert the minimum possible communalities in the diagonal cells is to make the assumption at once that there is only one common factor and that each test has a very large specific factor, and to make this assumption solely on the mathematical grounds that the low rank can be attained, or very nearly attained, without any psychological consideration whether these large, these enormously large, specific factors actually exist.

It seems to me that we are only justified in using minimum communalities if (*a*) we know that the correlations are only due to a certain small number of common causes, as was the case in Burt's experiment, or (*b*) we think it worth while to make the assumption of a similar number of hypothetical common causes for convenience of thinking. In psychological experiments we can seldom, if ever, know that there is actually only a small number of common causes acting. It may be worth while to speak and to act " as if " this were the case, but what I am afraid of is that these purely hypothetical entities may be held to have a proved existence in a much more definite sense, sometimes even in a physical sense.

REFERENCES

1. Thomson, Godfrey H. (1916). *Brit. J. Psych.*, VIII, 271–81.
2. Thomson, Godfrey H. (1919). *Proc. Roy. Soc. Lond.*, A, XCV, 400–8.
3. Thomson, Godfrey H. (1927). *Brit. J. Psych.*, XVIII, 68–76.
4. Mackie, John (1928). *Brit. J. Psych.*, XIX, 65–76.
5. Mackie, John (1929). *Proc. Roy. Soc. Edin.*, XLIX, 16–37.
6. Thomson, Godfrey H. (1934). *Nature*, CXXXIV, 700.
7. Thomson, Godfrey H. (1935). *Brit. J. Psych.* XXVI, 63–92.
8. Thomson, Godfrey H. (1949). *The Advancement of Science*, VI, 23, 267–74.

A REPLY TO SIR GODFREY THOMSON'S NOTE

I am most grateful to Sir Godfrey Thomson for his comments on my paper. With his warning, and with the gist of his final conclusion, I am, I think, in general agreement. He puts his finger on what to my mind is the most vexed question at the present moment, namely, are psychologists justified in substituting 'reduced correlations' for unity in the diagonal of the matrix to be factorized? He thinks this change unjustifiable. Presumably he would revert to Pearson's procedure—though he does not explicitly say so, nor can I recollect any factorial investigation where he has himself adopted that method.

Space does not allow me to consider his arguments in detail. But I should like to make the following brief comments.

1. In the foregoing paper I was only concerned to compare methods of weighting, not to adduce evidence for the assumption of "one factor causing the correlations." I have nowhere attempted to prove such a conclusion by "reversing the argument"; and I fully agree with Thomson that it could not be done that way. If I did not insert a more emphatic caution on this point, that was simply because I had dealt with this particular criticism in the last number of this *Journal*.[1]

2. My reduction of the 'self-correlations' is based not, as Thomson seems to suppose, on considerations of rank, but on considerations of significance. (I avoided the term 'communality' just because that has become associated with the idea of reducing rank.) My general grounds for preferring reduced self-correlations to self-correlations of unity have been given in earlier papers, and were summarized on p. 113. With these grounds, however, Thomson does not deal. The object of "the experiments with lifted weights, etc.," was, as stated in the opening sentences of Section IV, merely "to illustrate the effects of different types of weighting" where an objective criterion was available, not to justify the reduction of self-correlations or communalities as a general principle. That would have been far too wide a question to re-examine in an article dealing with another problem.

3. Thomson goes on to say that "in the experiment with lifted weights we *know* that the correlations are entirely due to one factor." With this, I fear, I cannot agree. On the contrary, I strongly suspect that the correlations are due to more than one common factor. My point was, not that no further factors existed, but that, in the experiment described, the additional factors were too small to be statistically significant; and in such a case, so I should maintain, all we can take for granted is the existence of the one common factor which is established as fully significant.

4. Nor can I agree that this result was "a foregone conclusion." In fact the outcome of the experiment on judging areas seems to show that it was not. Here there *were* other factors.

5. At the same time, in thus assuming that (within the margin allowed by the probable errors) the correlations can be accounted for by means of only one common factor, I am not for one moment supposing that this single factor is therefore simple. The fact that, as Thomson suggests, a general factor like intelligence may be biologically due to a large number of genes, or neurologically due to a large number of neural bonds, does not appear to me to be in the least incompatible with its operation as a single independent factor. Sweden may be a single independent State though it consists of over six million persons.

6. Thomson would prefer to say that in cases such as I have described "the tests have acted 'as if' they had only one factor in common." With this form of expression I should heartily agree. But then, surely every empirical generalization in any and every science can state no more than this. For example, in mechanics, chemistry, and thermodynamics observable changes take place 'as if' energy was conserved; but does anyone suppose that energy and similar "purely hypothetical entities" have a "proved existence in any more definite sense"?

7. Thomson, like other critics, may have misunderstood my insistence, here and in previous papers, on the inclusion of an external criterion (such as the teacher's ranking for 'general intelligence') and on the calculation of the correlation between this criterion and the hypothetical factor-measurements. Some have apparently inferred that the object of such calculations was

---

[1] 'The Two-Factor Theory,' this *Journal*, II, p. 152. With the criticisms advanced in his most recent paper (8) I have already dealt in *Brit. J. Sociol.*, I, pp. 162–4. My agreement with Thomson's general view is already implied in the text, where I deprecated the common confusion between 'factors as statistical components' and 'factors as concrete entities'—"a confusion" (I added) "against which Professor Thomson has so rightly and so often protested" (p. 106): cf. also *Factors of the Mind*, pp. 213f., Section on 'The Reality of Factors.' If I am asked, how do I attempt to establish the existence of 'general ability,' I should reply, as I have so often said, that the hypothesis is based on many converging lines of concrete (non-mathematical) evidence; and I regard factor analysis merely as a way of verifying that hypothesis. Thus the argument is formally, not deductive, but *inductive*, or, as it is now the fashion to say, 'hypothetico-deductive': (cf. *Brit. J. Sociol.*, p. 162, footnote 3 and refs.). Thomson, I gather, is inclined to reject the inference: as he has said elsewhere, it is (in his view) "not necessary to postulate the existence of a general factor," and he is "not personally a believer in the existence of a faculty called 'general ability.'"

literally to identify the general factor with a 'known common cause' (as Thomson calls it); and they have accordingly asked: why not go further, and *maximize* the correlation of your factor with that criterion? My reply[1] would be that, with a properly planned experiment, the general or basic factor should be far more trustworthy than the external criterion. The calculation of the correlation between the two was intended to throw light on *both*—to correct the empirical criterion as well as to check the interpretation and labelling of the factor. Thus, in the case of 'general ability' or 'intelligence' (which, I take it, is the kind of 'hypothetical common cause' Thomson has more especially in mind) I should regard the factor-measurements derived from 'a psychological set of correlations between tests' as furnishing better estimates than the 'criterion' provided by the teacher, though his assessments might at first sight seem more objective. May I add that in my view the 'assumption of such a hypothetical common cause' is made, not merely 'for convenience of *thinking*,' but also for the convenience of *practical* work—e.g., allocation to schools or predictions of future performances?

CYRIL BURT.

ERRATA

1. On p. 29 of the previous issue (Vol III, Part I) in Table IV A, col. 3 (bipolar factor III), the signs for the first 13 traits should be negative, and for the next 7 positive (the division thus corresponding with the division between group factors E and G); the next 9 should also be positive, and the last 9 negative (to agree with the allocation of signs adopted on p. 32).

2. On p. 67 in Table XVIII B, the figures printed in the first two rows and columns correspond with an alternative division of the first four tests (namely, into 1–2 and 3–4). To correspond with the division adopted in XVIII A and C (into 1 and 2–4), Table XVIII B should therefore read as follows:

	1	2–4	5–7	8–10
1	·7805	1·5695	1·3995	1·5105
2–4	1·5695	3·4206	2·7605	2·9795
5–7	1·3995	2·7605	3·1996	2·7200
8–10	1·5105	2·9795	2·7200	4·9601

---

[1] I should also answer that, as a rule, it would be better to maximize the correlation of the criterion with the *tests* than with the *factor* (i.e., to use ordinary multiple correlation): but here I am concerned only with aims, not with methods. I may add that, when I first suggested the device of correlating internal factors with external criteria, I used *two alternative criteria*: the object was to see with which of these two abilities the general factor could best be identified. And I went on to emphasize that the factor itself, "though single is not simple, but complex" (4, 1917, p. 55).

THE VALIDATION OF SEASHORE'S MEASURES OF MUSICAL TALENT BY FACTORIAL METHODS

By JOHN McLEISH

Extra-Mural Department, University of Leeds

I. *Problem.* II. *Subjects and Tests.* III. *Results :* (a) *Factor Analysis of the Seashore Battery ;* (b) *Multiple Correlations with External Criteria.* IV. *Summary and Conclusions.*

I. PROBLEM

The object of the research described in the following paper was to attempt to validate the Seashore tests of musical ability by a factorial procedure, using both internal and external criteria. The need for both kinds of validation appears to be now generally admitted in the more familiar fields of intelligence and educational testing. But it has been almost wholly ignored in the construction of tests for musical ability.

The problem of scientific validation is one of the most difficult with which the constructor of any kind of test has to deal. The commonest practice is to take some external criterion with which the test-results are then compared. This, however, is logically unsatisfactory, since it is just because external evidence is tacitly assumed to be unreliable that we have undertaken to construct a standardized test. The argument on which the validation is based is consequently circular : a criterion, admittedly imperfect, is used to demonstrate the worth of the test which is to replace it. The use of multiple criteria does not altogether release us from the logical dilemma to which we are thus reduced, for the multiplication of dubious or biased testimony does not necessarily make the conclusion more certain.

The argument may perhaps best be illustrated from the field of intelligence testing. In his earlier work, Spearman used as his criterion the ratings for intelligence submitted by teachers. He believed that the unreliability of their judgments could be eliminated by securing multiple criteria, i.e., ratings from two or more teachers for the same children, and then correcting the raw correlation between the average judgments of the teacher and the various tests he used. But, as Burt showed in his 1909 article,[1] the judgments of most teachers have a bias towards verbal ability and ability to learn ; he therefore proposed instead to use an *internal* criterion in the shape of the 'highest common factor' underlying the whole table of correlations between the tests selected according to a psychological plan. At the same time, in deciding the nature of the internal factor so found, as he went on to insist, it was not sufficient to trust to the names of the tests or to subjective notions of what they are intended to test ; an objective check, based on improved *external* evidence, is always desirable. The process is then, not a vicious circle, but rather a spiral of successive approximation.

The danger of omitting this double check becomes especially clear in the endeavours made from time to time to validate tests of musical ability, such as the Seashore battery (9). Dr. Stanton's 'Eastman experiment' is perhaps the most comprehensive and painstaking attempt at validation of any musical test so far carried out. The methods used in that experiment (11) and in other investigations by the same author (12, 13) included retests of students of music after training, comparison of their scores in the tests with their teachers' ratings, the development of a cumulative 'key for prognosis' by empirical studies, a comparison of the gradings furnished by this key with all other data available about the students—their class-marks, recitals, dismissals, etc. On the basis of the results, Dr. Stanton concludes that the 'cumulative key' (which included both the Seashore records and an intelligence test) forms a "useful administrative tool for the selection and classification of students ; . . . the tests, which became the basis of classification, can do quickly what the faculty could learn only after long observation and experience with the students" (11).

At the same time she fully recognizes that the manifestation of musical talent depends on many circumstances apart from the possession of the innate ability (or set of abilities) which the Seashore tests were designed to measure. Among the more obvious of the extraneous influences that play an important part in determining the level of performance or appreciation, she enumerates the following: "poor health, insufficient finances, lack of ambition or effort, family demands, personality problems, religious conflicts, 'inferiority complexes,' insufficient emotional poise, misdirected goals, disintegrated concepts of prerogatives [*sic*], incoherent plans, unfavourable rapport between teacher and pupil, poor musical training, inadequate musical preparation for the educational level attempted, extreme personal jealousies, marriage, over-anxious parents, exploitation by patrons, teachers, or parents" (11, p. 131). Such a list is, in effect, a list of flaws in the normal validating procedure.

---

[1] 'Experimental Tests of General Intelligence,' *Brit. J. Psychol.*, III (1909), pp. 168f.

The investigation reported in the following pages should be regarded as a supplement to Dr. Stanton's investigations rather than a criticism of them. Its chief aim was to make good the more obvious shortcomings of previous enquiries, namely, the failure to apply available statistical techniques and in particular the techniques of factor analysis.

The earliest mention of 'factors' underlying musical ability, which is based on any formal statistical procedure, appears to be that contained in the Board of Education's *Report on Psychological Tests of Educable Capacity* (introductory chapter, prepared at the request of the Committee by Burt, p. 20); and in his survey of group factors, published a year or two later (2), he mentions "a distinct musical capacity, no doubt highly complex," as definitely established. The data on which this conclusion was based were obtained from experiments carried out by Burt and Pelling for purposes of educational and vocational guidance, partly in L.C.C. schools and partly at the National Institute of Industrial Psychology. The novel feature in the approach consists in the addition of a number of more complex test-items, intended to measure "the Gestalt-qualities of musical understanding, as contrasted with the atomistic approach adopted by Seashore and most other investigators hitherto."[1] Attempts were later made by a succession of research-students, working at the Psychological Laboratory, University College (17 and refs.), to determine the more important components of this 'highly complex' capacity: the factors found included elementary perceptual abilities both auditory and kinæsthetic, auditory and kinæsthetic imagery, a broad factor for æsthetic appreciation generally, a more specialized factor for musical appreciation as such (largely influenced, it would seem, both by musical experience and by temperamental inclinations), and finally general intelligence.[2] The last and the most recent of these researches was that carried out by Wing (18: for fuller details reference should be made to his long and elaborate thesis (20), which also gives a fairly complete historical review).

Several writers have published the inter-correlations between the tests in the Seashore battery; Brown (1) and Drake (4) have published figures which seem fairly representative; and Wing (18, 19) reports the results of a factor analysis of the Seashore tests by Burt. With the exception of the last, which is admittedly very tentative in its nature and limited in its scope, none of those who have attempted to assess the validity and reliability of the battery have taken the logical step of factorizing the tables of correlations between its constituent tests. Instead, both Seashore's disciples and his critics seem often content to rest their conclusions largely on anecdotes of personal impressions. As a result, it is impossible to reconcile the views of American investigators like Stanton with the contrary statements of critics like Wing, who contends that the tests are atomistic, boring, musically meaningless, irrelevant, and invalid (21).

The methods of factor analysis would seem to be pre-eminently adapted for the solution of the logical difficulties which arise in studying such batteries. It is unfortunate, however, that the objectivity which results from the use of such techniques is so often lost by the way in which the factors are subsequently identified. This, as we have seen, often rests solely on subjective interpretation. Even those who have claimed to demonstrate a factor of musical ability underlying their tests have either used no external check at all or else have been satisfied to compare the test-results with ratings obtained from teachers, whose opportunities for judging genuine musical ability are, as a rule, extremely limited: ability in playing the piano or singing in tune is hardly a safe criterion.

---

[1] Burt, C., 'The Psychology of Music' (Gresham Lectures, 1932, unpublished, cited by Wing (18), p. 83); also *id.*, 'The Factorial Analysis of Emotional Traits,' *Char. and Pers.*, VII, 1939, pp. 293–9.

[2] 17, pp. 266f., and refs. The tests used by Burt and his collaborators consisted, not only of tests of the more elementary processes—discrimination of pitch, rhythm, loudness, etc.—but also of appreciation of melody, harmony, short musical passages, longer extracts, and 'altered versions' (given partly by the piano and partly by gramophone records). Paper-and-pencil tests were also included, e.g., marking well-known tunes in order of preference for a concert programme.

In the light of the criticisms expressed by Burt and his students—Wing, Williams, and others—it was felt that a new approach to the question of validation of the battery was needed. In particular, Seashore's assertions that intelligence and other non-musical factors had little or no influence on these measures of 'elemental' musical ability required to be examined in the light of experimental results. Part of the aim of the present investigation therefore was to attempt to establish whether the factors resulting from an analysis of the correlations between the several tests could be approached more objectively by comparing the factors with independent evidence regarding other musical and non-musical abilities of the persons tested. As a result it is claimed that a closer approximation to an objective interpretation has been achieved.

The question that presents itself at the outset may be formulated as follows. What elements in the testing situation, apart from differences in musical abilities of the subjects, might conceivably affect the scores obtained? A preliminary study of the test battery and the relevant literature suggested that the following elements were probably important (8, 10):

(i) *Age*. In view of Seashore's expressed opinion about his 'measures of musical talent'—that pitch discrimination (8, p. 65), sense of time (p. 113), consonance (p. 157), tonal memory (p. 242), intensity (p. 99), and even the sense of rhythm (p. 126) are innate *elemental* capacities not affected to any considerable extent by training or environment—it is evidently desirable to ascertain the actual relations existing between the age of the subjects tested and their scores on the six tests in the battery. Seashore holds that these six capacities mature quickly in very early childhood; so that by the time the child sufficiently understands the instructions it is possible to obtain a quantitative measurement of his natural capacities. On the other hand, Stanton and Koerth (12) have found marked differences between the performances of pre-adolescent, adolescent, and post-adolescent groups, which obviously throws doubt on Seashore's assertion. They explain the discrepancy as arising from the fact that, in the case of the adult, one is actually measuring the *psycho-physical* threshold, whereas with the child one is measuring the *cognitive* threshold. The former, they hold, cannot be measured until the age of maturity, which they place at sixteen years. Unfortunately they report no correlations between age and score for the various groups studied.

(ii) *Intelligence*. Seashore, as we have seen, claims that intelligence has very little influence on the scores obtained with his battery, provided, of course, that the subject possesses the necessary minimum of intelligence required to obey the instructions (8, pp. 56, 98, 113, 157). On the other hand, Burt (quoted by Wing, 21, p. 342) found that "when intelligence has been partialled out, the net correlations between the tests drops to little over 0·10, and are thus, with the groups we have tested, statistically insignificant." He concludes that most of the correlation between the tests, at any rate with children of the type he tested (namely, pupils at elementary schools), is due quite as much to intelligence as to innate musical ability. With older groups, more homogeneous in general intelligence, for example, children of a secondary (grammar) school type or students, the effect of intelligence proved to be somewhat less important.

(iii) *Speed of Judgment*. On introspective grounds, speed of judgment would seem to be an influence that might have quite a considerable weight in affecting the scores. With the tests of Memory, Rhythm, and Consonance, although the stimuli are presented at a speed suitable for distinct hearing, nevertheless the test itself requires a definite decision: it is a matter, not of just listening, but also of making up one's mind about small points of difference; and at the rate the stimuli are given mere quickness of decision would seem to be important.

(iv) *Interest and Boredom*. As previously remarked, it has been suggested by Burt and Wing (21) that, for persons who are interested in music, the lack of a genuinely *musical* content renders the Seashore tests comparatively meaningless and boring. This, they say, becomes most conspicuous with the professional musician. Seashore noted much the same reaction in certain musicians he tested; however, on retesting, after taking care to establish better rapport and eliminate the attitude of negativism, the difficulty, he claims, was overcome, and the correct psycho-physical level of the subjects secured (10).

(v) *Musical Training*. Seashore (8, 10) (and both Stanton and Koerth confirm his opinion (12, 13)) claims that the scores of subjects are uninfluenced by the amount of musical training they have had before the test. According to Seashore, the tests are truly elemental, that is, "so simple and natural that the resulting record does not vary with training" (8, p. 65).

In view of these and other considerations, the following problems were formulated for experimental study:—

(*a*) Does the Seashore battery measure a unitary ability, such as would be manifested by a general factor, or does it measure different and independent aspects of hearing, as Seashore himself suggests?

(*b*) Assuming that the battery does measure a unitary ability, is it possible to say in what this consists? Alternatively, if there is a 'general factor,' is it conceivable that it might represent not a unitary ability, but an aggregate of many common elements?

(*c*) In the latter case, assuming that these common elements can be statistically eliminated e.g., by partial correlation, and that there is still a residuum left over and above these, can this residuum be identified with 'musical talent'?

(*d*) Lastly, since Wing has been the most consistent critic of the Seashore tests, how far do the scores obtained with the Seashore battery by a given group of subjects agree with their scores for Wing's own tests?

It will be seen that this approach to the problem of validation is in direct contrast to Seashore's own attitude. His procedure can be described as basically *ad hoc*. He conceives of the 'measures' primarily as a sieve for the elimination of the musically unfit, rather than as a quantitative instrument for measuring musical ability. His use of the psychogram in place of weighted or unweighted averages for describing individual endowment is the natural corollary of his view that the abilities measured are independent. There seems in fact, as we have already implied, to be a striking similarity between Seashore's conceptions and methods and those of intelligence testers of the pre-factor era. Seashore's supporters might perhaps retort that the later development of a mathematical superstructure has tended to make its employers lose sight of significant individual peculiarities behind a mass of numerical coefficients. In Seashore's statistical methods one never forgets the individual; but, on the other hand, no clear psychological picture of the abilities tested is developed.

II. SUBJECTS AND TESTS

The tests described below were applied to a group of 100 subjects, all undergraduate or postgraduate students of psychology. To ensure comparable conditions in the procedure, every test was administered and scored by the writer. The experiments were treated as part of a normal laboratory course in experimental psychology, and in this way maximum co-operation was ensured. The series of experiments was presented as a research project, the reasons for each variable in the tests and methods being explained in advance, so as to arouse interest, but always of course so as not to interfere with the proper application of the tests. As a result, interest and motivation were very high. The two or three students who at first appeared uninterested were afterwards found to be 'tone-deaf' or to have an exceptionally low degree of musical sensitivity. In all, the following 22 variables were included in the enquiry.

1. *Age.*

2. *Intelligence* (measured by the Cattell Group Test IIIA, (*a*) timed (for 'speed'), (*b*) untimed (for 'power').

3. *Speed of reaction.* For the purpose of measuring this variable, a cancellation test was used. A reaction time test, with an auditory stimulus, was impracticable with so large a group.

4. *The Oregon test of musical discrimination.* This was scored in two parts: (*a*) correct preference for the original musical passage; (*b*) recognition of the nature of the alterations made in the original piece (i.e., whether the alteration is of melody, harmony, or rhythm).

5. *Recognition of orchestral instruments.* This test was included in order to ascertain the amount of informal musical training or interest possessed by each subject. It consisted of fifteen instruments playing in succession the same passage from Schubert's *Ave Maria*, the instruments being recorded on a Wirek recorder in random order. The subjects were asked to name the instrument, or, failing this, to say to which section of the orchestra it belonged.

6. *The Seashore measures of musical talent.* These consist of measures of pitch discrimination, intensity, time, rhythm, consonance, and tonal memory (9). Seashore's procedure was followed, except that in the case of pitch, intensity, and time, when the subject could detect no difference, he was asked to reply 'equal,' instead of guessing.

7. *Rating for interest or boredom.* Immediately after the application of each musical test, the subjects rated it for interest on a five-point scale (20). Interest was broadly described as the ease with

which attention could be maintained throughout the test, and the various points on the scale were carefully defined.

8. *Wing's test of musical intelligence.* This consists of the seven tests—chord analysis, pitch change, memory, rhythmic accent, harmony, intensity, phrasing—presented, like the Seashore battery, on gramophone records (21). Unfortunately, only half of the subjects completed all the seven tests in the battery ; the rest completed only the first five. Nearly all the subjects found some difficulty in sustaining attention, even though two sessions were allowed with rest pauses between. This perhaps is partly attributable to the recording, and partly to the use of the piano for every test.

9. *Questionnaire on musical interest and training.* A questionnaire of the open or essay type (15) was freely adapted from the questionnaire used by Seashore (8, p. 17). The subjects were encouraged to write at length on their interests in music, and to be as specific as possible in mentioning composers and pieces.

It should be emphasized that the methods adopted were a compromise between the possible and the desirable. In particular, the cancellation test, the interest ratings, and the questionnaire were included only as rough trials ; and the conclusions base on them must be regarded as suggestive rather than conclusive.

III. RESULTS

In this paper only those results which have a bearing on the validity of the Seashore battery will be reported. In the calculation of data, all scores were first normalized and standardized. Calculations were taken to four places of decimals, but only two places are printed.

(*a*) *Factor Analysis of the Seashore Battery.* The correlations obtained are shown in Table I. With one non-significant exception they are all positive ; and range from little over zero to 0·43. Except four (all correlations with Rhythm) every one is significant.

TABLE I. INTER-CORRELATIONS OF THE SEASHORE TESTS

	Pitch	Intensity	Consonance	Memory	Time	Rhythm
Pitch	(·85)	·39	·27	·42	·25	·14
Intensity ..	*·05*	(·89)	·18	·22	·28	·10
Consonance ..	*·06*	*·06*	(·34)	·43	·24	—·12
Memory ..	*·05*	*·06*	*·05*	(·93)	·23	·43
Time	*·06*	*·06*	*·06*	*·06*	(·81)	·03
Rhythm ..	*·06*	*·07*	*·07*	*·05*	*·07*	(·82)
Factor I ..	·62	·48	·39	·87	·42	·28
Factor II ..	·18	·27	·16	—·42	·26	—·45

Split-half reliabilities are shown in the diagonal cells, and probable errors below the diagonal.

The coefficients are of the same order of magnitude as those reported by Brown (1) ; while the reliabilities are somewhat higher than those reported by Brown or Drake (1, 4). The higher reliabilities are perhaps attributable, at least in part, to the use of the reply ' equals ' in cases of doubt.

A factor analysis was carried out by Burt's method of simple summation (3), using successive approximations for the self-correlations. The saturations are appended at the foot of the table. Since, in addition to the factor-saturations so obtained, we have figures for the unreliability of each test, the results of the analysis may be expressed in terms of Burt's four-factor theory (3, p. 103 and refs.). We have in fact factors of four different types : (i) a general factor, (ii) a bipolar factor, (iii) a number of specific factors peculiar to each test, and (iv) factors of error or unreliability.

(i) *The General Factor*. The general factor accounts for about 30 per cent. of the total variance. The question of greatest importance is the psychological nature of this general factor.

As has already been indicated, the common practice is to interpret each factor by a mere scrutiny of the tests into which it enters. This practice has been sufficiently criticized by Burt in this *Journal* and elsewhere ; and would certainly be most inappropriate here. Fortunately we have been able to secure results from a number of other tests, musical and non-musical, selected, in accordance with the general plan of the research, as likely to throw light on the problem. Let us begin by comparing the saturations obtained by the six tests of the Seashore battery with their correlations with the various musical and non-musical tests. The result is a suggestive series of figures, set out in Tables II and III.

TABLE II. SATURATIONS FOR FIRST FACTOR AND CORRELATIONS WITH MUSICAL TESTS

Seashore Test	Factor I	Wing	Oregon (*a*)	Oregon (*b*)	Question-naire	Orchestral Instruments
Memory ..	·87	·68	·67	·73	·33	·26
Pitch	·62	·61	·47	·47	·24	·18
Intensity ..	·48	·22	·27	·41	·23	·30
Time	·42	·25	·34	·23	·03	·12
Consonance ..	·39	·44	·46	·43	·36	·14
Rhythm ..	·28	·31	·31	·45	·01	·18
Average ..	·51	·42	·42	·45	·20	·20

The Seashore tests have been rearranged according to their first factor-saturations. If we compare this order with the orders indicated by their correlations with the musical tests, the correspondence is so close as to provide a strong confirmation of the hypothesis that the general factor underlying the Seashore battery is intimately related to musical capacity. If we ignore the test of Consonance (and the justification for this will emerge in the sequel), it will be seen that the questionnaire gives exactly the same order ; Wing's tests and the Oregon test (*b*) (nature of alteration) correspond just as closely except for Rhythm ; the Oregon test (*a*) (piece preferred) corresponds except for Intensity. On the other hand, knowledge of the instruments of the orchestra test shows little correspondence.

Table III enables us to compare the order of the factor-saturations with that of the correlations obtained with the non-musical tests.

TABLE III. FIRST FACTOR-SATURATIONS AND THE NON-MUSICAL TESTS

Seashore Test	Factor I	Cattell (timed)	Cattell (untimed)	Age	Interest	Cancella-tion
Memory ..	·87	·32	·18	−·05	·34	·15
Pitch	·62	·22	·19	·02	·39	·00
Intensity ..	·48	·13	·25	−·03	·40	·02
Time	·42	·13	·15	·17	·06	−·03
Consonance ..	·39	·03	·00	−·00	·23	·12
Rhythm ..	·28	·17	·19	−·17	−·05	·16
Average ..	·51	·17	·16	·07	·25	·08

The coefficients themselves are now much lower. There appears to be some relation, presumably causal, between the six Seashore saturations and the ratings for ' intelligence ' (especially ' timed ') and for ' interest ' ; but none with age or ' cancellation.' It would seem that speed at the low level measured by the cancellation test is not particularly important for any of the tests, but that speed at the higher levels measured by the Cattell test exercises an appreciable influence on the tests of Memory and Pitch.

The conclusions to be drawn from Tables II and III may be summarized as follows. (i) The general factor in the Seashore battery is related to musical ability on the appreciation side (as measured by the Oregon test) and possibly on the interest and performance side (as measured by the questionnaire). It is similar to the ability measured by the Wing test. (ii) It is, however, also influenced by intelligence, particularly when the test employed takes the form of a speed test.

The next step is to check the analysis by removing the influence of the timed intelligence score from the inter-correlations of the six Seashore tests. This was carried out in the usual way by the aid of Yule's formula for partial correlation. The results are shown in Table IV. A comparison of Tables I and IV shows that, even after ' timed intelligence ' has been removed, the pattern of relations remains much the same as before.

TABLE IV. SECOND-ORDER CORRELATIONS : TIMED-INTELLIGENCE PARTIALLED OUT

Test	Pitch	Intensity	Consonance	Memory	Time	Rhythm
Pitch	—	·36	·27	·36	·23	·10
Intensity ..	·36	—	·18	·18	·26	·08
Consonance ..	·27	·18	—	·43	·24	—·12
Memory ..	·36	·18	·43	—	·19	·39
Time	·23	·26	·24	·19	—	·01
Rhythm ..	·10	·08	—·12	·39	·01	—

(ii) *The Bipolar Factor*. The bipolar factor can be more briefly dismissed.

The saturations are hardly at all affected by partialling out intelligence. On comparing the saturations with the same external characteristics as before (see Table V), only in the case of age and the cancellation test can any appreciable relation be discerned.

TABLE V. THE BIPOLAR FACTOR AND THE NON-MUSICAL TESTS

Seashore Test	Factor II	Cattell (timed)	Cattell (untimed)	Age	Interest	Cancellation
Rhythm ..	—·45	·17	·19	—·17	—·05	·16
Memory ..	—·42	·32	·18	—·05	·34	·15
Consonance ..	·16	·03	·00	·00	·23	·12
Pitch	·18	·22	·19	·02	·39	·00
Time	·26	·13	·15	·17	·06	—·03
Intensity ..	·27	·13	·25	—·03	·40	·02
Average * ..	·29	·17	·16	·07	·25	·08

* Calculated regardless of sign.

The bipolar factor is best interpreted as a factor of classification. It appears to divide the Seashore tests into three subgroups : as will be seen from column 1, the first two tests have high negative saturations, the next two non-significant positive saturations, the last two much higher positive correlations, which are probably quite significant. The first two comprise Rhythm and Memory : here the tests involve a *sequence* of taps in the case of Rhythm, and of tones in the case of Memory. Since the material seems so obviously meaningful, these two tests are commonly selected by musicians (21) as the best in the battery. However, their saturations with the musical factor do not altogether bear out this verdict: for, although Memory yields the highest figure, Rhythm has the lowest. The second group includes Pitch and Consonance. These are tests involving an immediate comparison of elementary stimuli, in both cases musical tones or combinations of tones. Next to Memory these two tests had the highest correlations with other musical tests (Table II). The third group includes Time and Intensity. Here the task again requires an immediate comparison of elementary stimuli—clicks in the one case and the sounds of a buzzer in the other. This group therefore appears to depend chiefly on stimuli of a non-musical nature. The results might perhaps be tentatively summed up by saying that the bipolar factor appears to contrast tests involving the immediate perception of a change with those involving the immediate memory of a change.

TABLE VI. THE BIPOLAR FACTOR AND THE MUSICAL TESTS

Seashore Test	Factor II	Oregon (*a*)	Oregon (*b*)	Orchestral Instruments	Wing	Question-naire
Rhythm	—·45	·31	·45	·18	·31	·01
Memory	—·42	·67	·73	·26	·68	·33
Consonance	·16	·46	·43	·14	·44	·35
Pitch	·18	·47	·47	·18	·61	·24
Time	·26	·34	·23	·12	·25	·03
Intensity	·27	·27	·41	·30	·22	·23
Average	·29	·42	·45	·20	·42	·20

(iii) *Error Factors.* The proportion of the variance attributable to unreliability or 'error factors' is rather high. It accounts for more than 20 per cent. of the total variance. Nearly half of this is contributed by the test of Consonance (split-half reliability only 0·34).

The test of Consonance is notoriously difficult to apply, because of the preconceptions on the part of most subjects. The judgments of subjects musically untrained appear to rest on a mixture of vague impressions of 'harmonious combination,' general feelings of pleasantness or unpleasantness, and perhaps other emotional states and associations (6, 7). As Burt and others have pointed out, the test is plainly based on an old-fashioned conception of harmony, which has largely become out of date as a result of the development of modern music (5, 7, 20, 21). In the introspective reports on the test, many of the comments alluded to the influence of context and *Gestalt*, whereby even a rough combination of notes can be accepted as consonant, provided that the discord appears to be satisfactorily resolved either mentally or by the succeeding combination. This was stressed by at least a dozen subjects. A small number of subjects explicitly mentioned the effect of a familiarity with modern harmony on their judgments of relative dissonance. Evidently these students were not strictly obeying the instructions given, namely, to compare the roughness or smoothness of the two combinations themselves.

It may be added that a supplementary analysis was undertaken by examining the partial correlations obtained after intelligence, age, and performance in the cancellation test had been eliminated. However, the elimination of the further variables appeared to leave the general pattern of relations unaffected; and it is therefore unnecessary to report the figures in detail.

(iv) *The Specific Factors.* The factors discussed so far, namely, the general factor, the bipolar factor, and the factors of unreliability, account for only 62 per cent. of the total variance. The remaining 38 per cent. must therefore be explained by specific factors, i.e., by factors which are more or less peculiar to each of the six Seashore tests.

The specific factors may perhaps [1] be interpreted as indicating the amount of the 'atomism' in the tests, so that a comparison with tests of musical ability, and, in particular, with those devised by Wing, should throw considerable light on the extent to which this criticism is justified. With Wing's tests a second bipolar factor, almost as large as the first, was obtained; and in this case both bipolar factors appeared to be fully significant. Hence only about 12 per cent. of the variance

[1] The objection of 'atomism,' urged by Burt and others against the Seashore battery, does not turn solely on the fact that the tests measure highly specific processes; it also implies that they measure only the more elementary constituents (which, of course, tend also to be highly specific). Such processes, so these writers argue, are necessary but not sufficient. Their contention is that "musical appreciation as such is concerned far more with the complex Gestalt-like qualities of musical composition than with the mere discrimination and memory of pitch, intensity, basic rhythms, etc."

in Wing's battery remains attributable to specific factors. Wing's battery, therefore, is not quite as atomistic as Seashore's. But if, as most factorists seem agreed, musical ability is, in Burt's phrase, " highly complex," some degree of specificity or atomism would seem desirable. As has often been stated, " the most economical type of test-battery is one in which all the tests have a high correlation with the general factor, but for the rest have a low correlation with each other."

Final Analysis. One of the aims of the present investigation was to compare the Seashore battery with that of Wing. In all the Wing tests used in the present research the piano is the source of the stimuli ; and passages or chords with some musical significance are employed. A factorial analysis of the scores obtained by the same group of subjects with the seven tests devised by the Wing battery shows that, with his tests, the general factor accounts for as much as 45 per cent. of the total variance ; and, as we shall see in a moment, it is possible, by an appropriate weighting of the Seashore tests, to secure a multiple correlation of 0·72 between the two batteries. It may therefore be concluded that Wing's tests measure much the same kind of ability as Seashore's, but measure it at a higher or at least a different level, namely, that of musical meaning. The results of the entire analysis for both batteries are shown in Table VII.

TABLE VII. A COMPARISON OF THE FACTORIZATION OF THE WING AND SEASHORE TESTS

The figures indicate the proportion (expressed as a percentage) contributed by each factor to the total variance.

Tests	General Musical Ability	First Bipolar	Second Bipolar	Specifics	Unreliability
Wing	45	10	8	12	25
Seashore ..	29	10	0	38	23

TABLE VIII. FINAL ANALYSIS OF THE SEASHORE BATTERY

Tests	Factors				
	G(*m*)	*G*(*i*)	*B*	*Sp*	*E*
Memory	·76	·32	·31	·39	·26
Pitch	·59	·22	—·17	·65	·29
Intensity	·46	·13	—·26	·77	·33
Time	·39	·03	—·24	·76	·44
Consonance	·43	·13	—·16	·37	·81
Rhythm..	·27	·17	·52	·62	·41
Percentage of Variance	25	4	10	38	23

The final analysis of the Seashore tests is set out in Table VIII. The first four columns are based on the analysis of the partial correlations in Table IV. *G*(*m*) denotes that part of the general factor attributable to musical ability as such ; *G*(*i*), that part attributable to intelligence ; *B*, the bipolar factor ; *Sp*, the specific factors ; and *E*, the factors of error or unreliability. From these figures a complete picture of the battery can be built up, enabling us to evaluate in precise factorial terms both the claims made on its behalf and the criticisms urged against its shortcomings.

(*b*) *Multiple Correlations with External Criteria.* The preceding results were obtained when each of the six tests of the battery was allowed the same weight. The conclusions drawn, however, show that the value of the several tests as indications of musical ability varies widely from one test to another. The Consonance test, for example, is apparently devoid both of validity and of reliability, and therefore should probably receive no weight whatever. Indeed, in his *Revised Measures of Musical Talent* (which unfortunately were not obtainable when the above experiments were planned) Seashore himself has dropped it. What we now want to know, before using the test for practical purposes, is (1) how would the correlations be altered if differential weights were employed in combining the scores, and (2) what would be the highest correlation obtainable, if we used the best possible weights ?

As regards weights, the requisite regression coefficients, calculated in the usual way, are shown in Table IX below.

TABLE IX. REGRESSION COEFFICIENTS FOR THE SEASHORE TESTS

Test	Pitch	Intensity	Consonance	Memory	Time	Rhythm
Oregon (*a*) ..	·17	·02	·21	·42	·14	·13
Oregon (*b*) ..	·12	·21	·21	·44	—·02	·25
Wing ..	·39	—·08	·18	·39	·04	·12
Questionnaire	·03	·08	·18	·20	·21	—·07
Interest ..	·20	·33	—·02	·33	—·14	—·25
Intelligence ..	·08	·19	—·09	·08	·07	·11
Age	·04	—·08	—·08	·01	·20	—·19

The multiple correlations are given in Table X, and reveal with greater clarity the relations between the Seashore battery as a whole and the other tests employed. The squares of these correlations indicate the proportion of the variance contributed by the item in question.

TABLE X. MULTIPLE CORRELATIONS AND VARIANCES

Test	Correlation	Variance
Oregon I ..	·75	·55
Oregon II ..	·82	·67
Wing	·72	·62
Questionnaire..	·48	·23
Interest ..	·45	·21
Intelligence ..	·32	·10
Age	—·03	·00

IV. SUMMARY AND CONCLUSIONS

1. Seashore's tests of musical ability were applied to 100 undergraduate and postgraduate students ; and an attempt was made to assess the validity of each component test both by internal criteria (i.e., determining the saturations of each test with the factors entering into all) and by external criteria (i.e., determining the correlation of each test with independent assessments obtained from other objective sources, more especially supplementary tests of musical and non-musical abilities).

2. The inter-correlations were factorized by Burt's method of simple summation ; and, since reliability coefficients can readily be calculated for the separate tests, the results eventually obtained may be conveniently expressed in terms of his ' four-factor type of hypothesis.' They reveal (i) a general factor accounting for 29 per cent. of the variance, (ii) a bipolar factor accounting for 10 per cent., (iii) specific factors accounting for 38 per cent., and (iv) error factors accounting for 23 per cent.

THE
BRITISH JOURNAL
OF
PSYCHOLOGY

Statistical Section

EDITED BY
CYRIL BURT AND GODFREY THOMSON

WITH THE ASSISTANCE OF
CHARLOTTE BANKS AND ARTHUR SUMMERFIELD

AND THE FOLLOWING EDITORIAL BOARD

A. C. AITKEN
M. S. BARTLETT
W. G. EMMETT
M. G. KENDALL
D. N. LAWLEY
E. S. PEARSON
L. S. PENROSE
J. FRASER ROBERTS
A. RODGER
B. BABINGTON SMITH
W. STEPHENSON
P. E. VERNON

Managing Editor : H. G. MAULE

Volume III
1950

UNIVERSITY OF LONDON PRESS LTD.
WARWICK SQUARE, LONDON, E.C.4.

3. The interpretation of the several factors was based partly on the introspective reports secured from the subjects on the mental processes used in performing each of the Seashore tests, but mainly on the external evidence obtained from the supplementary tests. The following conclusions were drawn :—

(i) The saturations for the general factor were first systematically compared with the correlations obtained between each test in the Seashore battery and the tests or assessments of both musical and non-musical characteristics. It was found (*a*) that age and (*b*) musical knowledge have little or no influence on the general factor ; (*c*) that intelligence (especially when assessed by speed of performance) has a small amount of influence ; and (*d*) that performances in the Wing and Oregon tests show the closest agreement. It may therefore be inferred that the general factor is closely related to musical ability and appreciation.

(ii) The bipolar factor is evidently a classification factor. It appears to subdivide the six Seashore tests according to their concrete content, contrasting (*a*) those that depend chiefly upon immediate discrimination with (*b*) those that depend chiefly upon immediate memory.

(iii) The specific factors for Pitch, Intensity, Time, and Rhythm are decidedly large ; those for Memory and Consonance much smaller. It is suggested that the saturations for these factors may be taken as indicating in some degree the ' atomistic ' nature of the battery, i.e., the extent to which it depends upon a number of specific and mutually independent abilities.

(iv) The unreliability of the battery as a whole appears to be rather high. But this is due mainly to the extreme unreliability of the test of Consonance.

4. On these and other grounds it is argued that the common criticisms of the Seashore battery, as expressed by Wing and others—namely, that it is invalid, irrelevant, excessively atomistic, and musically meaningless—are not altogether borne out by the results obtained. A comparison of the Seashore tests with those of Wing suggests that the two batteries measure much the same general musical factor, but at different levels and from rather different aspects. The lower validity of the Seashore battery seems due largely to its greater specificity, and might be easily improved by omitting the test of Consonance and modifying the speed at which the test-items are presented.

5. It is concluded that, in its general nature, the Seashore battery is adequate for its original purpose, namely, to measure the more elementary abilities required for the understanding and appreciation of music ; but that its use will be most effective if the scores are weighted in accordance with the calculated regression coefficients, and if it is used in conjunction with other tests of musical appreciation.

REFERENCES

1. Brown, A. W. (1928). ' The reliability and validity of the Seashore tests of musical ability.' *J. Appl. Psychol.*, XII, 468–76.
2. Burt, C. (1925). *The Measurement of Mental Capacities.* Edinburgh : Oliver and Boyd.
3. Burt, C. (1940). *The Factors of the Mind.* London : University of London Press.
4. Drake, R. M. (1933). ' The reliability and validity of tests of musical talent.' *J. Appl. Psychol.*, XVII, 447–58.
5. Lambert, C. (1948). *Music Ho !* Pelican Books Edition.
6. Pickford, R. W. (1948). ' The borderline of psychology, physics and music.' *Nature*, CLXI, 589.
7. Pickford, R. W. (1949). ' Experiments on the relation of dissonance and context.' *Quart. J. Exp. Psychol.*, I, 57–67 ; III, 107–18.
8. Seashore, C. E. (1919). *The Psychology of Musical Talent.* New York : Silver, Burdett and Co.
9. Seashore, C. E. (1919). *Measures of Musical Talent.* New York : Columbia Phonograph Co.
10. Seashore, C. E. (1938). *The Psychology of Music.* New York : McGraw-Hill.

11. Stanton, H. M. (1935). 'The measurement of musical talent; the Eastman experiment.' *Univ. Iowa Studies*, II.
12. Stanton, H. M., and Koerth, W. (1930). 'Musical capacity measures of adults repeated after musical education.' *Univ. Iowa Studies*, XXXI.
13. Stanton, H. M., and Koerth, W. (1933). 'Musical capacity measures of children repeated after musical training.' *Univ. Iowa Studies*, XLII.
14. Vernon, P. E. (1936). 'Tests in æsthetics.' *ap.* Hamley, H. R. *The Testing of Intelligence.* London: Evans Bros. n.d. 133–41.
15. Vernon, P. E. (1938). *The Assessment of Psychological Qualities by Verbal Methods.* London: H.M. Stationery Office.
16. Whipple, G. M. (1915). *A Manual of Physical and Mental Tests.* Baltimore: Warwick and York.
17. Williams, E. D., Winter, L., and Woods, J. M. (1938). 'Tests of literary appreciation.' *Brit. J. Educ. Psychol.*, VIII, 265-84.
18. Wing, H. D. (1936). *Tests of Musical Ability in Children.* (Unpublished Thesis: University of London Library.)
19. Wing, H. D. (1941). 'A factorial study of musical tests.' *Brit. J. Psychol.*, XXXI, 341–55.
20. Wing, H. D. (1941). *Musical Ability and Appreciation.* (Unpublished Thesis: University of London Library.)
21. Wing, H. D. (1948). 'Tests of musical ability and appreciation.' *Brit. J. Psych. Mon. Sup.*, XXVII.

NIGHT VISUAL CAPACITY AND ITS RELATION TO SURVIVAL IN OPERATIONAL FLYING

By D. D. REID

London School of Hygiene and Tropical Medicine

I. *Problem and Method.* II. *Results.* III. *Summary and Conclusions*

I. PROBLEM AND METHOD

In his Moynihan lectures, Livingston (1944) has described the development and use, in the Royal Air Force during the second world war, of his Rotating Hexagon Test for night visual capacity. Its introduction as a selective method was justified largely on *a priori* grounds, although early results suggested that success in the test was predictive of success in action. Thus the scores for night visual capacity obtained by a group of successful night-fighter pilots were appreciably above the average of flying personnel in general (Livingston, 1948).

The reliability of the test, and its relevance to the performance of tasks of identification in conditions of low illumination, has been fairly thoroughly studied (Bradford Hill and G. O. Williams, 1943). But without clearer indications of the operational importance of testing for night vision, some degree of opposition to the introduction of the Hexagon Test as a method of selecting or excluding men hoping to go forward for flying training was almost inevitable. Moreover, with the increased use of radar aids in aerial gunnery and in landing, ability to see at night became less important during the later stages of the war. Nevertheless, it continues to be relevant both to flying and to other occupations in peace time ; and this would seem to justify the publication of the following results obtained during an attempt at securing an operational validation of the Hexagon Test.

The results of the Hexagon Test of night visual capacity, which was usually applied at an early stage in training, were recorded on each man's medical documents ; these accompanied him throughout his service and were available to the squadron medical officer. In the course of an investigation already reported (Reid, 1947), the records of all the operational crews on active service in Bomber Command of the Royal Air Force were examined and the N.V.C. (Night Visual Capacity) scores were noted, together with the number of bombing sorties completed by the individual by the date of the survey. Then, by producing frequency distributions of N.V.C. scores for men who had completed 0, 1, 2, 3, or more sorties up to the usual tour limit of 30 sorties, it was easy to calculate the mean or average N.V.C. score among groups of men at different stages of their operational tour. If night visual capacity was important in determining the length of survival, i.e., if a selective elimination of those with poor night visual capacity took place by their being shot down early in their tour, then this should be revealed by a progressive rise in the average values of the N.V.C. scores among those who survive to complete most or all of the tour expected of them. If, on the other hand, ability to see at night was quite unimportant in determining survival, the level of the mean N.V.C. scores in groups of increasing lengths of operational experience should remain, within sampling limits, unchanged.

It is clear that night vision is unlikely to be equally important to the wireless operator, whose duty never allows him to use that faculty at all, as it is to the gunners, upon whose speedy appreciation of an impending attack by an enemy night-fighter the safety of the aircraft may well depend. Separate tabulations are thus needed for each type of crew duty.

In view of the many factors of personal skill, mechanical efficiency and all the hazards of weather and war which are involved in a bomber crew's chance of survival, it would be idle to expect any very close relation between an individual's night vision test score and the number of sorties he survived. Consistent rather than spectacular trends must be looked for ; and, even where these trends are clear enough, the comparison of such trends in the various types of crew duty demands the use

TABLE I. MEAN N.V.C. SCORE AT DIFFERENT STAGES OF TOUR

Duty	Number of Sorties completed to date													
	0		1–4		5–8		9–12		13–18		19–30		Total	
	No.	Mean	No.	Mean	No.	Mean	No.	Mean	No.	Mean	No.	Mean	No.	Mean
Pilots	65	18·5	132	19·9	111	18·8	89	19·9	106	19·3	119	20·9	622	19·6
Navigators	94	18·6	123	18·5	100	18·1	91	19·1	102	19·7	112	18·9	622	18·8
Bomb-Aimers	91	18·7	138	18·2	119	17·6	108	19·4	102	19·8	123	19·9	681	18·9
Wireless Operators	114	18·3	192	18·6	148	20·3	111	19·1	143	19·2	146	19·1	854	19·1
Flight Engineers	91	17·5	82	18·3	55	18·0	30	16·9	43	18·7	51	19·9	352	18·2
Mid-upper Gunners	131	18·8	192	18·9	166	19·0	136	19·4	150	19·6	164	20·1	939	19·3
Rear Gunners	149	19·8	195	19·8	164	19·5	134	19·5	166	20·4	178	20·0	986	19·9
Total	735	18·7	1,054	19·0	863	18·9	699	19·3	812	19·7	893	19·9	5,056	19·3

of some statistical method of describing them, and of testing whether apparent differences are larger than could reasonably be due to chance. For these reasons, regression lines have been fitted to the data for each type of crew duty to describe the trend, if any, of the mean night visual capacity score in groups of men of differing operational experience. The significance, in the statistical sense, of these trends and of the differences between them can then be readily assessed by the appropriate analysis of variance technique.

II. RESULTS

The N.V.C. scores were grouped according to the number of bomber operational sorties carried out by the individual up to the time of the survey (0, 1–4, 5–8, etc.) to give series of roughly equal size. The mean N.V.C. scores of the men in each of these 'experience-groups,' divided according to the type of aircrew duty performed, are set out in Table I.

In Table I the totals certainly show no very dramatic rise in mean N.V.C. score in groups of successively longer operational experience. In navigators and wireless operators no clear trend is apparent, but for mid-upper gunners, on the other hand, there is a consistent rise in mean score with each increase in the number of sorties carried out by the date of the survey. To study and compare these differences more precisely, the correlation coefficient between N.V.C. score and operational experience measured in individual sorties was calculated and a straight line fitted to the data to represent the change in mean N.V.C. score with each increase in the number of sorties survived. The regression coefficients (N.V.C. on number of sorties) whose sizes indicate the slopes of this line for each crew duty are set out in Table II.

The Total row in Table II shows that there is a very small, but quite definitely significant, positive relationship between the N.V.C. score at initial examination and the number of sorties survived up to the time of the survey. Further, the apparent progressive rise in mean N.V.C. score with increasing operational experience, such as could have occurred by chance less than once in a thousand times, is best represented by a line whose slope is such that the mean score rises by ·0516 units between groups of men with 1, 2, 3, etc. sorties, to their credit.

TABLE II. RELATION BETWEEN N.V.C. SCORE AND OPERATIONAL EXPERIENCE

Duty	Number	Correlation Coefficient r	Regression Coefficient b	Critical Ratios t
Pilot	622	·0825	·0646	2·0604*
Navigator	622	·0365	·0267	0·9106
Bomb-Aimer	681	·1350	·0939	3·5502**
Wireless Operator	854	·0358	·0268	1·0459
Flight Engineer	352	·1125	·0927	2·1185*
Mid-upper Gunner	939	·0878	·0593	2·6992**
Rear Gunner	986	·0260	·0172	0·8163
Total	5,056	·0715	·0516	5·0974**

* Significant : $P < \cdot 05$ ** Highly significant : $P < \cdot 01$

As Table III shows, this trend can be adequately described by a straight line since the residual variability left after fitting it differs insignificantly from the 'error' (in this case the within-array) variance. The same holds for the trends of N.V.C. score in successive groups of increasing operational experience in the specific aircrew duties where the trend is significantly upwards. The regression lines for the four main groups, based on the figures in Table I, are drawn in Fig. 1.

As the tests of significance in Table II suggest, the trends of these straight lines are upward for pilots, bomb-aimers, flight engineers and mid-upper gunners. For wireless operators and navigators there is no such consistent trend : in other words, there is no evidence of a possible selective elimination in action of wireless operators or navigators with below average N.V.C. scores. This, of course, is what might have been expected, since in neither of these aircrew duties is night vision important.

TABLE III. ANALYSIS OF VARIANCE

Source	Sum of Squares	D.F.	Mean Square	F
PILOTS				
Residual	872·3152	29	30·0798	< 1
Regression	175·7040	1	175·7040	4·19**
Between arrays	1,048·0192	30	34·9340	
Within arrays	24,794·9392	591	41·9542	
Total	25,842·9584	621	—	—
NAVIGATORS				
Residual	773·3008	29	26·6655	< 1
Regression	29·9280	1	29·9280	< 1
Between arrays	803·2288	30	26·7743	
Within arrays	22,606·4688	591	36·5592	
Total	23,409·6976	621	—	—
BOMB-AIMERS				
Residual	1,208·8896	29	41·6858	1·36
Regression	391·0592	1	391·0592	12·80**
Between arrays	1,599·9488	30	53·3316	
Within arrays	19,855·4640	650	30·5469	
Total	21,455·4128	680	—	—
WIRELESS OPERATORS				
Residual	937·2624	29	32·3194	< 1
Regression	39·7536	1	39·7536	1·09
Between arrays	977·0160	30	32·5672	
Within arrays	30,031·2272	823	36·4899	
Total	31,008·2432	853	—	—
FLIGHT ENGINEERS				
Residual	1,215·4224	29	41·9111	< 1
Regression	205·7056	1	205·7056	4·45**
Between arrays	1,421·1280	30	47·3709	
Within arrays	14,830·3264	321	46·2001	
Total	16,251·4544	351	—	—
MID-UPPER GUNNERS				
Residual	1,010·5504	29	34·8466	1·20
Regression	212·5984	1	212·5984	7·23**
Between arrays	1,223·1488	30	40·7716	
Within arrays	26,344·5280	908	30·3609	
Total	27,567·6768	938	—	—
REAR GUNNERS				
Residual	909·5888	29	31·3651	1·01
Regression	19·8752	1	19·8752	< 1
Between arrays	929·4640	30	30·9821	
Within arrays	28,434·4144	955	29·7743	
Total	29,363·8784	985	—	—
TOTAL				
Residual	929·6608	29	32·0573	< 1
Regression	895·0416	1	895·0416	26·00**
Between arrays	1,824·7024	30	60·8234	
Within arrays	173,136·5120	5,025	34·4550	
Total	174,961·2144	5,055	—	—

** Highly significant : P < ·01

Conversely, for mid-upper gunners, the implied importance of night visual capacity agrees with reasonable expectation since their part in 'spotting' enemy fighters might well be decisive for the aircraft's defence by gunnery or evasive manœuvre. Similarly, as Bradford Hill and Williams have shown (1943a), the pilot may require good night vision for landings at night under operational conditions. It is conceivable, too, that the bomb-aimer and flight engineer may from their alternative positions in the front gun cockpit and astradome play a part in the 'spotting' of enemy aircraft which would explain the relevance of night vision to their survival; but the negative findings for rear gunners makes such an explanation less convincing.

As Table II shows, there is no significant upward trend in N.V.C. for the rear gunner category. For this there seem to be two possible explanations. First, it may be that night vision is unimportant in the defensive role played by the rear gunner; but the positive inferences noted for mid-upper gunners render this unlikely. Alternatively, there remains the possibility that the initial selection of men with superior night vision for employment as rear gunners (implied by their higher mean N.V.C. score before commencing the tour) did, in fact, eliminate men whose defective night vision might have proved a danger to the aircraft in which they flew.

TABLE IV. ANALYSIS OF VARIANCE

Source	Sum of Squares	D.F.	Mean Square	F
Deviations from own 'duty' regression	172,824·6992	5,042	34·2770	—
Between 'duty' regressions ..	179·5808	6	29·9301	—
Deviations from common regression	173,004·2800	5,048	34·2718	—
Common regression	895·0416	1	895·0416	26·12**
Within duties	173,899·3216	5,049	34·4423	—
Between duties	1,061·8928	6	176·9823	5·14**
Total	174,961·2144	5,055	—	—

** Highly significant: $P < \cdot 01$

The complete analysis of variance for the data is set out in Table IV. From this table it appears that the differences between the mean score in each crew category are significantly large—a reflection of the policy already noted of selecting those with better results on the Hexagon Test for aircrew duties which, on *a priori* grounds, appeared to demand superior night vision. Then, looking within the various types of duty, part of the differences between the individual scores recorded is clearly accounted for by the regression, common to all duties, between night vision and the number of sorties completed. The variance ratio test shows that the mean square due to this regression is significantly greater than the estimate of random or 'error' variation given by the deviations of individual observations from this common regression line.

The regression line describing the upward trend of mean N.V.C. score with increasing experience may, of course, have a different slope among groups of men with differing crew duties. Indeed, the slope of such trends might be taken to indicate the relative importance of night visual capacity in determining the chance of survival in different duties. A steep slope (such as that shown in Fig. 1 for bomb aimers and flight engineers) would imply that the upward trends in mean N.V.C. score are due to a high rate of elimination of men with inadequate night visual capacity.

It would be tempting to speculate on the possible explanations for the suggestive disparities between the slopes among mid-upper gunners and flight engineers. That such speculation would be premature, however, is evident from the top section of Table IV where the variability between these slopes is shown to lie within the limits of 'error' variation. On the whole, this overall analysis amplifies and confirms the deductions already drawn. It seems evident that in certain types of duty, i.e., heavy bomber pilots, bomb-aimers, flight engineers, and mid-upper gunners, night visual capacity as measured by the Hexagon Test was significantly, if slightly, related to the time of survival in operational flying. The general course of the upward trend in N.V.C. score seemed to hold for all four types of duty since the differences between them were insignificant, and there was no evidence that this trend was other than linear. On the other hand, although the absence of trend among 'inside' duties (navigators and wireless operators) agreed with reasonable expectation, it was also absent for rear gunners for whom night vision had been considered to be of prime importance. However, as suggested above, this last feature of the data seems in all probability to be the result of selection. There remains the question—is the general correlation between Hexagon score and duration of survival really the expression of a direct causative relationship between night visual capacity and operational skill?

For the phenomena observed, several explanations may be advanced. It might be that the upward trend in N.V.C. score with increasing operational experience was the result of a relaxation of standards laid down at initial selection. Such a relaxation would presumably result in the more recent recruits to a squadron's strength having a lower average N.V.C. score than their more experienced colleagues. Yet not only is there no record of any such change in policy having taken place, but also such a change would have produced a fairly sudden change in the level of the N.V.C. score rather than the consistent linear trend actually observed. Again, a change such as the complete neglect of N.V.C. testing in selection in general policy would have been expected to have some effect on the trend in *all* types of duty. In fact, as already noted, the trend was conspicuously absent among the 'inside' members of the bomber crews.

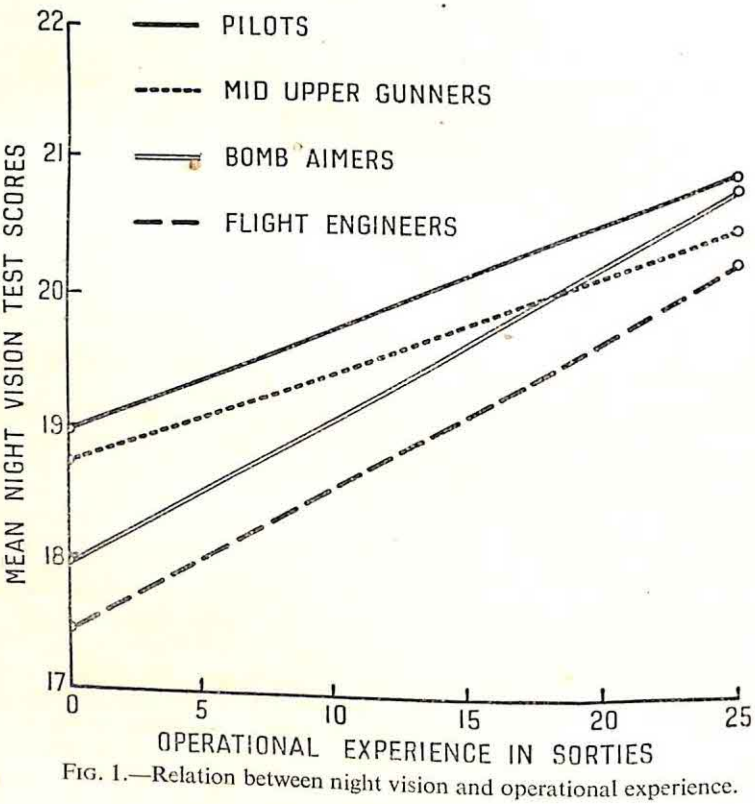

FIG. 1.—Relation between night vision and operational experience.

Yet another explanation would be that the correlation between Hexagon Test score and operational survival depends not on any direct cause and effect link but on some common relation to a third factor, presumably some personal quality or ability which is really effective in determining an individual's chance of completing his operational tour. Intelligence and neurotic predisposition may be two such characteristics. If, for example, intelligence, as measured by the usual paper and pencil tests is positively related both to the results of the Hexagon Test and to the number of sorties likely to be survived, then any selective elimination in action of the less intelligent would result in a simultaneous reduction in the number of men with below average Hexagon Test scores. Similarly, there is some evidence that neurotic predisposition as assessed by psychiatric interview is related to operational or flying efficiency (Reid, 1942, Davis 1948). There is also evidence which suggests that patients suffering from neurotic illness do rather poorly at the Hexagon Test (Livingston, 1945). It does not follow from this, of course, that these men would have done poorly in the test *before* the onset of their illness, e.g., at initial examination before exposure to operational stress—their performance may be a sign of their acute illness rather than a permanent feature of their native abilities. The possibility that there is some such underlying explanation for the results observed must, however, be investigated.

Some light on the problems involved may be obtained by examining the results of intelligence tests, night vision capacity tests, and psychiatric interviews conducted in the course of another investigation on a large series of pilots under training. Considering the nature of the Hexagon Test, it would not be surprising to find that some of the factors involved in intelligence tests of a conventional kind might have an influence on the results achieved. As it happened, the intelligence tests administered in this particular experiment were divided into a test of form perception (V.C.), a test of verbal intelligence (I.M.A.) and a test of mathematical reasoning (C.M.). As Table V shows, none of these sub-tests shows any significant relation to the results attained by the same individuals in the Hexagon Test. Again, results of a general test of intelligence gave similar negative findings on correlation with Hexagon Test scores (Parry, personal communication). It follows, then, that the correlation of Hexagon score with operational survival does not depend upon any coincident correlation with intelligence in so far as intelligence is measured by the tests used.

TABLE V. CORRELATION OF HEXAGON TEST AND INTELLIGENCE TESTS RESULTS

Test	r	n	t
V.C. . .	·0735	367	1·4077
I.M.A. . .	·0009	367	·0000
C.M. . .	·0419	365	·8091

Note : all coefficients are insignificantly different from zero.

In the other investigation mentioned (Symonds and Williams, 1948) neurotic predisposition was assessed by two psychiatrists who, on the basis of a biographical interview, classified the men examined into four categories of predisposition to the risk of breakdown—nil, slight, moderate, and severe. The N.V.C. scores given by the Hexagon Test for the individuals thus classified were then divided up according to the usual conventions—Excellent (30–32), above average (20–29), average (9–19), below average (3–8) and poor (0–2). Contingency tables, summarizing the relations existing between broad groupings of these categories both of psychiatric assessment and Hexagon Test results, are given separately for each psychiatrist in Table VI.

For both psychiatrists, although the trend in results is suggestive, there is no significant relationship between assessment and N.V.C. score : (for Psychiatrist A, $\chi^2 = 2{\cdot}3877$, $n = 2$, $\cdot 50 > P > \cdot 30$; for B, $\chi^2 = 1{\cdot}1791$, $n = 2$, $\cdot 70 > P > \cdot 50$; together, $\chi^2 = 3{\cdot}5668$, $n = 4$, $\cdot 50 > P > \cdot 30$). In so far as these results give a fair picture of this relation, then there is no reason to suppose that the observed trend in mean N.V.C. score with increasing operational experience is really dependent on the part, if any, played by neurotic predisposition as a factor in survival. The very slight relation appearing in the table may be enough to explain the insignificant rise in N.V.C. score which takes place in all duties with increasing experience.

To sum up, the evidence suggests that the significant upward trend in mean N.V.C. score found in all but one of the ‘ outside ’ types of bomber-crew duty is an expression of some causal relation between the abilities detected by the Hexagon Test and the number of sorties survived, i.e., that night visual capacity, irrespective of any association with either intelligence or neurotic predisposition, is of definite importance in operational flying.

As the low correlation coefficients between N.V.C. score and length of survival suggest, the importance of night vision, relative to all the other factors involved in surviving the many-sided hazards of operational flying, is comparatively small. Indeed, less than 1 per cent. of the total variance in length of survival is accounted for by differences in night visual capacity as measured by the Hexagon Test. But this is hardly surprising. Quite apart from the many factors involved in operational survival, there is a close interdependence between all members of a crew; and this combined with the fact that the Hexagon Test was used to eliminate those with more severe deficiencies in night vision, must tend appreciably to dilute the correlation between test result and length of survival. It is indeed a little unexpected that any such correlation could be demonstrated in operation data of this kind.

TABLE VI. RELATION BETWEEN PSYCHIATRIC ASSESSMENT AND N.V.C. SCORE

	N.V.C. Score		
Psychiatrist A's Assessment	Above Average (8–32)	Below Average (0–7)	Total
Nil	34 (48·57%)	36 (51·43%)	70
Slight	32 (42·10%)	44 (57·90%)	76
Mod. and severe	7 (30·43%)	16 (69·57%)	23
	73	96	169
Psychiatrist B's Assessment	Above Average (8–32)	Below Average (0–7)	Total
Nil	32 (39·02%)	50 (60·98%)	82
Slight	23 (33·82%)	45 (66·18%)	68
Mod. and severe	6 (27·27%)	16 (72·73%)	22
	61	111	172

The operational validation of tests of all sorts in the Services or in industry is expensive in terms of time, labour, and money. Theoretically, validation requires the initial testing of all subjects whose subsequent career in the chosen pursuits is then followed up. But in circumstances where new tests are being introduced, there is probably no entirely satisfactory substitute for this approach. Where the need is for a rapid assessment of some personal fact (e.g., a physical measurement or biographical detail), some variant of the method used in this study may be useful, particularly if only current personal records are available.

III. SUMMARY AND CONCLUSIONS

1. In the hope of assessing the importance of night visual capacity as measured by the Livingston Hexagon Test in the determination of operational survival, a statistical analysis was made of the records of over 5,000 men in Bomber Command of the Royal Air Force in 1944. The score achieved by each man tested *before* starting his operational tour was noted, together with the number of sorties survived up to the date of the survey.

2. The mean N.V.C. scores among men who had completed 0, 1, 2, etc., up to 30 sorties were then computed and the presence of an upward trend in these means which might be the result of the selective elimination in action of those with lower N.V.C. scores was tested for significance. Significant upward trends suggestive of a positive correlation between N.V.C. score and the length of operational survival were found among three of the 'outside' members of a crew (pilots, bomb-aimers, mid-upper gunners) and for flight engineers who had occasional look-out duties, but not for rear gunners and the two 'inside' members of the crew, viz., navigators and wireless operators.

3. The association noted in these cases does not appear to be the result of any correlation between night visual capacity and intelligence or between night visual capacity and intelligence and neurotic predisposition. There was no demonstrable relation between the scores obtained with the Hexagon and the results of testing of intelligence or neurotic predisposition. The unexpected result in rear gunners may be explained by the rigorous selection standards on the Hexagon Test demanded for rear gunners.

4. The general conclusion is that, whatever its precise mechanism, night visual capacity as measured by the Hexagon Test had a definite, if small, influence on survival in operational flying.[1]

REFERENCES

1. Bradford Hill, A., and Williams, G. O. (1943). 'Statistical studies of night visual capacity as measured by the rotating hexagon test.' *Flying Personnel Research Committee Report*, No. 533.
2. Davis, D. R. (1948). 'Pilot error.' *Air Ministry Air Publication*, 3139A. H.M. Stationery Office.
3. Livingston, P. C. (1944). 'Visual problems of aerial warfare.' *Lancet*, II, 33–8.
4. Livingston, P. C. (1945). 'Practical values and clinical applications of night vision.' *Med. Press*, CCXIII, 4–10.
5. Livingston, P. C. (1948). 'Experiences in night vision in the Royal Air Force.' *Transactions of the Ophthalmological Society of Australia*. VI, 44–58.
6. Reid, D. D. (1942). 'Influence of psychological disorder on efficiency in operational flying.' *Flying Personnel Research Committee Report*, No. 508. H.M. Stationery Office.
7. Reid, D. D. (1947). 'Some measures of the effect of operational stress on bomber crews.' *Air Ministry Publication*, 3139, Chap. XIX. H.M. Stationery Office.
8. Symonds, C. P., and Williams, D. J. (1948). 'Psychological disorders in flying personnel.' *Air Ministry Publication*, 3139, Chap. XVI, 205–27. H.M. Stationery Office.

[1] I wish to thank the Director-General of Medical Services of the Royal Air Force, Air Marshal P. C. Livingston, both for his help and for his permission to publish this paper. In its preparation I have had the advice of the Royal Air Force Consultant in Medical Statistics, Professor A. Bradford Hill. To Mrs. K. M. Bull, Miss B. M. Miller, and Miss O. M. Penfold, I am indebted for computing and secretarial assistance, and to Mrs. M. G. Young for drawing the diagrams.

PRIMARY FACTORS OF PERSONALITY

By H. A. REYBURN and M. J. RAATH
The University of Cape Town

I. *The Experiment.* II. *The Centroid Analysis.* III. *The Orthogonal Analysis.* IV. *The Oblique Analysis :* (a) *Orthogonal Projection ;* (b) *Parallel Projection.* V. *The Factors.* VI. *The Interrelations of the Factors.* VII. *Summary.*

I. THE EXPERIMENT

The purpose of the experiment discussed in this paper was to reconsider and extend the results of factorial analyses of personality qualities already made by us (9, 10, 13). A larger range of qualities has been dealt with, the choice being governed partly by clinical experience and partly by analyses published by others, particularly Cattell (5).

There appear to be three main methods of dealing with the axes which a centroid analysis yields.

First, the axes, as they stand, may be taken to represent intelligible factors. The resulting general factor and subordinate bipolar factors are then regarded as concordant with the demands of logical definition (1). Against this must be set an important and perhaps decisive objection. The factors thus reached are relative to the battery of tests or variables, and thus lack objectivity (11, p. 397, and 12, p. 55).

Secondly, the axes may be rotated to find simple structure. For reasons given elsewhere we are not satisfied with this solution (12, 14) ; and in practice it does not seem to give the objectivity claimed. Indeed, the minimum conditions of simple structure are not always rigidly insisted upon, and the criteria for the supposed unique result turn out, in fact, to be a matter of degree (16, pp. 335 and 355).

Thirdly, factors may be identified positively and axes drawn through guiding points in the light of all available knowledge (12, pp. 59f. ; cf. 15, pp. 287f.). The criterion is then the intelligibility of the analysis as a whole and its concordance with other results. Certainty and objectivity are not to be expected at first. Factors which will remain objective in all relevant circumstances must be sought gradually and by a method of approximation. It is our hope that this further paper may contribute to this process.

For the present enquiry two groups of students (37 men and 46 women) acted as observers ; all had some acquaintance with psychology, and most were advanced students. Each was asked to select two subjects, approximately of his own age, and thoroughly well known to the observer. The total number amounted to 160, viz., 78 men and 82 women. About half were persons outside the University, and the sexes were fairly evenly balanced. In a few cases two observers chose the same subject : the correlations between their ratings were calculated as an indication of the reliability, and averaged ·806. The observers were first asked to sketch the main personality features of the two subjects chosen ; these were recorded under pseudonyms, and were remarkable for their freedom and candour. A questionnaire was then filled up, which required a number of qualities to be rated on a 5-point scale.

In Cattell's analysis bipolar qualities were frequently used : it was apparently assumed that psychologically, if not logically, the opposites were contradictories, not mere contraries. We were unwilling to make this assumption. Accordingly, the negative was taken to be the absence or extreme infrequency of the quality. It was left to the analysis itself to show if contraries could be treated as contradictories.

In administering the questionnaire the two groups of observers were taken separately. Each observer was given a typed copy of the questions and one of us went over the items with the whole group, giving further explanations. Each item was answered by everyone before the next question was considered; and stress was laid on the need of keeping the answers independent. There was one difference between the two groups. Item 26 deals with the level of ideal which the subject demands. The first group answered the question in a general way; the second divided it into two, reporting separately on the standard which the subject exacts from himself and on that which he exacts from others.

The ratings for each pair of qualities were correlated, first for each of the two groups of 80 subjects, and then for the full 160. The results served as a check on the reliability. In addition the answers were scrutinized in the light of the sketches already made.

II. THE CENTROID ANALYSIS

In making an analysis by simple summation or the so-called centroid method there is no infallible criterion to determine the number of factors to be extracted, and in any given experiment more may be present than the accuracy of the data allows one to extract. Weight must be given to the probability at each stage that the residuals can be explained as chance variations. This criterion was adopted by Reyburn and Taylor in analysing some of Webb's data (13, p. 157); but theoretical and practical considerations have since induced us to think that the analysis should be carried slightly beyond the stage thus indicated. In general it is desirable to bring the average residual below the probable error of a zero correlation; and attention should be paid to the rate at which the average contribution to the variance diminishes.

In the present experiment, after five factors had been extracted, the average residual was ·051; the P.E. of a zero correlation for 160 cases would be ± ·053. One more factor may perhaps be extracted. When this is done, the average residual falls to ·046. The average amount of variance extracted by each successive factor is (in percentages) as follows: Factor I, 14·8; II, 10·5; III, 10·9; IV, 5·4; V, 3·2; VI, 2·2. The saturations obtained for the several traits are shown in Table I. These may be called F_0.

The test of a successful rotation is the intelligibility of the factors obtained. Each should be a unitary function, and not a mere aggregate of accidentally associated qualities. It is difficult to rotate successfully if nothing is known beforehand of any of the factors. If, however, it is known that certain factors are present, the task is simplified, and it may be possible to establish new factors. In the method we adopt it is thus necessary to begin with some hypothesis regarding at least a number of factors, and the first and immediate test of its adequacy is the intelligibility of the set of factors as a whole. The nth factor must be as intelligible a unity as the first.

No single experiment, however, can finally establish the objectivity of any factors. Objectivity becomes finally manifest only when we can assume a set of factors in a variety of contexts and find on each occasion that the remaining factors also prove significant. The ultimate test is the coherence and intelligibility of the factors through the widest range of experimental material. In searching for factors which are objective in this sense, one must not expect to find them with high precision at the outset. If some variable is known to have a heavy load on a factor, it can be used as a guide to determine the best position of the axis. But the guide will not be perfect. The final position for the axes can only be approached gradually and in a series of approximations. In making an analysis, therefore, it is important to use past experience as well as present data.

The first decision to be made is whether the rotated axes are to be orthogonal or oblique. There are obvious advantages in favour of the former: the system is simpler, and the independence of the factors is a considerable asset. With the latter the angles between the axes are a part of the total picture; and whether they are resolved into factors of the second order or not, they add to the complexity of the analysis, and thus seem to offend against the principle of parsimony.

Orthogonal factors, however, are apt to be abstract and difficult to identify. Often they seem to consist in what is left of certain variables when what is due to other factors has been set aside. And although this presents no mathematical difficulty, it often presents psychological difficulties. Further, the order in which the factors are established is of the greatest importance. The possible load of the variables on any factor is limited by what has already been taken out of them. Thus the first factor tends to absorb the maximum loads of the variables, and the last one is only allowed to have what is left after all the correlations with the other factors have been deducted. These objections may not be fatal, but they are formidable. And they induced us to make a double analysis, using both orthogonal and oblique axes.

III. THE ORTHOGONAL ANALYSIS

We begin by fixing provisionally points through which the axes may be passed. To determine these points we made a survey of the data. The correlations were classified according to the significance which could be attached to them.

TABLE I. CENTROID FACTORS

No.	Trait	I	II	III	IV	V	VI	h^2
1.	Avoids Problems	·390	·278	—·575	—·200	—·038	·189	·638
2.	Overbold	·336	—·252	·323	·104	—·061	·069	·300
3.	Unsystematic	·373	—·084	—·380	—·289	·230	·128	·443
4.	Spontaneous..	·445	—·529	—·067	·111	·082	·101	·512
5.	Emotional	·331	—·211	—·450	·441	—·113	—·182	·597
6.	Emotionally variable ..	·328	—·240	—·093	·027	·110	—·092	·195
7.	Emotionally uncontrolled ..	·632	—·039	—·071	·261	—·122	·005	·490
8.	Relaxes easily	·270	—·148	—·224	—·428	·228	·008	·380
9.	Looks on bright side ..	·211	—·518	—·030	·000	—·065	—·093	·326
10.	Avoids strong emotions ..	—·503	·323	·086	—·090	—·091	—·118	·395
11.	Easily distressed	·142	—·048	—·263	·457	—·188	—·065	·341
12.	Prone to sad thoughts ..	·064	·415	—·259	·397	·057	—·058	·408
13.	Fits of depression	·206	·337	—·295	·303	·158	—·273	·434
14.	Avoids tests	·377	·302	—·387	—·198	—·151	·000	·445
15.	Feels inferior	—·050	·420	—·456	·243	—·032	·391	·601
16.	Shy	—·341	·417	—·396	—·018	—·065	—·120	·467
17.	Assertive	·327	—·118	·680	·070	·144	·000	·609
18.	Submissive	—·309	—·144	—·570	—·125	—·060	—·155	·485
19.	Subjective outlook	·329	·185	·362	·198	—·100	—·114	·336
20.	Fixed in outlook	—·395	·157	·209	·120	—·064	·306	·337
21.	Suggestible	·138	—·071	—·560	—·149	—·153	·201	·424
22.	Contra-suggestible	·252	·326	·380	·115	·175	·060	·362
23.	Incompliant	·311	·338	·432	·020	·275	—·253	·538
24.	Dominating	·480	·119	·472	·125	·022	·261	·551
25.	Self-critical	—·378	—·180	—·222	·158	·001	·172	·279
26*a*.	Realistic : Self	·443	·163	—·155	—·255	·666	·062	·759
26*b*.	Realistic : Others	·561	·208	—·179	—·367	·370	—·178	·693
27.	Aggressive	·527	·311	·335	·194	·072	·021	·530
28.	Withdrawn	—·339	·553	·098	·054	·000	—·072	·439
29.	Self-esteeming	·255	—·182	·598	—·136	—·055	—·175	·508
30.	Conceited	·510	·224	·300	—·202	—·314	·157	·564
31.	Self-reliant	—·374	—·173	·488	·203	·226	·003	·501
32.	Reliable	—·556	—·040	·264	·293	·142	·059	·490
33.	Persevering	—·490	—·159	·315	·274	·226	—·267	·563
34.	Cheerful	·360	—·700	·115	—·130	·038	—·053	·654
35.	Irritable	·455	·349	·036	·271	—·038	—·156	·429
36.	Excitable	·565	—·284	—·229	·250	—·150	·031	·538
37.	Easily rattled	·485	·111	—·258	·105	—·216	—·065	·377
38.	Suspicious	·214	·448	·235	·184	·089	·228	·396
39.	Kind	—·329	—·469	—·202	·106	—·101	—·104	·402
40.	Sensitive to others	—·478	—·308	—·225	·273	·064	·087	·461
41.	Touchy	·122	·310	—·258	·544	—·147	—·074	·500
42.	Buoyant	·001	—·616	·029	—·268	—·091	—·003	·460
43.	Blames others	·585	·313	·050	·042	—·236	—·038	·501
44.	Unbalanced	·426	·282	—·386	—·132	—·168	·091	·465
45.	Mixes badly	—·372	·585	—·166	—·029	·058	—·133	·530

In the first class were placed all which, in a sample of 160, could be expected to be significant at or below the ·001 level; in the second class those at or below the ·01 level; in the third, those at or below the ·02 level; in the fourth, those at or below the ·03 level; in the fifth, those at or below the ·05 level; in the sixth, those at or below the ·10 level; in the seventh class, those at or below the

·20 level ; and all the rest were regarded as devoid of significance. In the analysis attention was given almost exclusively to the first two classes.

The First Factor. One of the factors isolated from Webb's study on character and intelligence was named ' cleverness ' by Garnett (7) and ' surgency ' by Cattell (3 and 4).

In 1933 and 1936 Cattell indicated the close interrelationship between surgency and a number of other qualities. The relation was set forth as a bipolar one. This is shown in the following list which indicates the more intimate connections. *Positive Traits :* cheerful ; natural ; sociable ; humorous ; adaptable ; gregarious ; quick ; hasty. *Negative Traits :* gloomy ; formal ; unsociable ; earnest ; conservative ; exclusive ; slow ; introspective.

In 1939 Reyburn and Taylor found that the following qualities had significant factor loads on an axis placed to measure surgency-desurgency. *Positive Loads :* fondness for large social gatherings ; sense of humour ; cheerfulness ; corporate spirit ; bodily activity in pursuit of pleasure. *Negative Loads :* fits of depression ; conscientiousness. In 1941 Reyburn and Taylor dealt with the factor from the desurgency end, and found the following loads. *Positive Loads :* Indulges in self-pity ; loses head in excitement ; worries ; is easily hurt ; prefers working with others. *Negative Loads :* Takes prominent part in social affairs. In 1943 Reyburn and Taylor obtained the following results for surgency : *Positive Loads :* carefree ; happy-go-lucky ; unconcerned about others' opinions ; likes excitement ; unconcerned about future ; also talkative. *Negative Loads :* over-conscientious ; worries ; easily distressed ; thinks before acting ; dislikes being interrupted.

In 1944 Cattell made a fresh analysis of the ' Personality Sphere ' (4), and obtained a classification of traits for the bipolar factor (re-named surgency versus agitated, melancholic desurgency), which will be found in his book (5). A third and slightly different arrangement was presented in 1947 (6).

A consideration of these various analyses shows that there are several aspects to the activity which is being treated as a functional unity. On the one hand, there is an emotional aspect ; on the other, a more directly conative one, which may be called responsiveness. In our experiment, variable No. 4 (spontaneous) corresponds closely to the latter aspect ; the former one seems adequately covered by variables Nos. 9, 12, 13, 34, and 42. Their inter-correlations merit examination ; and are shown in Table II.

TABLE II. CORRELATIONS BETWEEN VARIABLES CLASSED AS RESPONSIVENESS

No.	Variable	4	9	12	13	34
9.	Looks on bright side ..	·423				
12.	Prone to sad thoughts..	−·156	−·280			
13.	Has fits of depression ..	−·038	−·237	·492		
34.	Cheerful..	·539	·510	−·418	−·399	
42.	Buoyant..	·349	·389	−·434	−·315	·595

The variables fall into two groups : Nos. 9, 34, and 42 present the surgent aspect ; Nos. 12 and 13, the desurgent aspect. The bond between the two sections is fairly strong, but weaker than within themselves. Moreover, they show a different relation to No. 4. Its correlation with the group consisting of Nos. 9, 34, and 42 is ·536 ; whereas with the two desurgent variables, Nos. 12 and 13, it is only −·123 : i.e., the two emotional aspects, though related, do not move together as a functional unity. Spontaneity carries cheerfulness with it, but does not exclude depression.

These considerations suggest that it is inadvisable to attempt, at least at the beginning, to locate a bipolar surgency-desurgency factor, and that it is better to use spontaneity as a guide to determine the positive aspect of the main factor in question. A trial of two arrangements, one in which spontaneity alone was used as a guide, and another in which it was combined with the centroid of variables Nos. 9, 34, and 42, showed little difference in the results. Accordingly the simpler form was used, and an axis was passed through No. 4, the factor obtained being provisionally named Spontaneity.

The Second Factor. The second factor represents Stability or integration. To reach it, an axis has been passed through the centroid of variables Nos. 5 reflected (not emotional) ; 13 reflected (free from fits of depression) ; 20 (fixed in outlook) ; 32 (reliable) and 37 reflected (not easily rattled).

This factor is in line with two of Cattell's factors, viz., *C* and *G* (5, pp. 317, 326, 480, and 489) and the main qualities attributed by him to them are reasonably well covered by the relevant relationships of our second factor as set forth in Tables V and VII below.

Although the positive aspect of surgency is largely accounted for by the first factor, the negative aspect remains almost intact. A desurgent factor, named '*d*' (mental depression), was described by the Guilfords in 1939 (8). But such a factor, when tried out on our data, although intelligible in itself, does not combine well with the factors that follow it, and in particular it makes the final two difficult to interpret. The various traits in the questionnaire which have a neurotic tinge are not reducible to a single factor; nor, if desurgency is abstracted from them, is the remainder readily intelligible. Moreover, neurotic deviations from the normal presuppose some positive principle of integration.

The Third Factor. The remaining factors may be described more briefly. The third factor is derived from Webb's conception of *w* (17). In 1939 Reyburn and Taylor, examining Webb's material, altered the conception chiefly by giving more weight to steadiness, continuity, and perseverance than to action from principle or purpose (9, p. 162). In 1943 they found the same factor appearing, although less clearly, in some of the Guilfords' material (13). A factor of this kind has also been postulated by other writers who have not set it in a factorial context. Variable No. 33 is taken as a guide to it here, and Persistence is suggested as a suitable name.

The Fourth Factor. One of Cattell's factors, *E*, is described as 'dominance (hypomania) versus submissiveness' (5, p. 321); and a similar quality has been recognized by others. It is represented fairly well in our list by variable No. 24 (dominating). But we have one alternative in No. 17 (assertive). As can be seen from the figures the two traits—dominating and assertive—agree in a good many points. But there are nearly as many significant failures to agree. Probably dominance is not a simple or ultimate unity, but involves assertiveness along with other elements. Eventually it was decided to pass an axis through No. 17. The factor of Assertiveness thus obtained resembles that of self-confidence as described by several other authors (18).

The Fifth and Sixth Factors. In the analysis of Webb's material (9, p. 162) Reyburn and Taylor found a factor which they named charity. In the list of qualities used in the present experiment there is a variable, viz., No. 39 (Kind), which might be expected to prove a guide to this factor if it is present in this analysis. Consequently an axis was passed through No. 39. The result, however, was unsatisfactory.

The fifth factor offers a choice, for it can be regarded from either end. On the one side, there are variables No. 11 (easily distressed), No. 36 (excitable), No. 5 (emotional), No. 41 (touchy), and on the other, No. 26 *a* and *b* (low standard of behaviour), No. 8 (relaxes), and No. 3 (unsystematic). The factor is bipolar, with sensitiveness, in perhaps more than one form or manifestation, at the one end, and toughness, callousness, and indifference at the other. The contrast of tender and tough presents the situation fairly well; and it seemed advisable to make Sensitiveness the positive end.

The sixth factor affords a measure of a feeling of Inferiority, and gives a clear, coherent picture.

The transformation matrix (T_1) which gives effect to these ideas is shown in Table III: the factor loadings resulting from this transformation may be called F_1.

TABLE III. ORTHOGONAL TRANSFORMATION MATRIX

·623	−·515	−·283	·511	−·052	·054
−·739	−·420	−·218	·314	−·044	·359
−·094	·571	·123	·785	·134	−·122
·155	−·105	·560	·008	·543	·597
·114	·161	·300	·058	−·813	·455
·141	·441	−·673	−·140	·150	·538

IV. THE OBLIQUE ANALYSIS

When oblique axes are used, there are two ways of dealing with the situation. On the one hand, each axis may be treated as if it were the first, no allowance being made for the correlation of the variables with the other factors; on the other, each factor may be treated as the last, and its loads calculated only after the effect due to the correlations of the variables with the other factors has been taken away. We may consider these two processes in turn.

(*a*) *First Method*. Using the same guiding points as before, we may pass an axis through each of them without reference to the position of the other axes. In doing this we have a choice of two starting points, *viz.*, the co-ordinates given by the centroid analysis itself, or those given by the orthogonal rotation. The use of the latter simplifies the arithmetic, and the transformation matrix required (T_2) is shown in Table IV below.

TABLE IV. OBLIQUE TRANSFORMATION MATRIX

1·000	−·447	−·249	·326	·242	−·271
0	·894	·519	·366	−·338	−·348
0	0	·817	·128	·342	−·348
0	0	0	·862	−·331	−·396
0	0	0	0	·774	·180
0	0	0	0	0	·704

The transformation matrix is triangular (*see* Burt, 1, p. 306); the factors obtained are oblique: and the cosines of the angles between them are given by the inner products of its columns, as follows:

TABLE V. CORRELATIONS BETWEEN FACTORS

Trait	1	2	3	4	5	6
1. Spontaneity.. ..	1·000					
2. Stability	−·447	1·000				
3. Persistence	−·249	·576	1·000			
4. Assertiveness ..	·326	·181	·213	1·000		
5. Sensitivity	·242	−·411	·044	−·287	1·000	
6. Inferiority	−·271	−·190	−·398	−·602	·203	1·000

(*b*) *Second Method.* If the axes obtained by the method just used are regarded as primary axes in Thurstone's sense of the term, the loadings obtained are the correlations of the variables with those axes; and F_2 will give the 'structure' of these factors. To find the 'pattern' for the same factors we must post-multiply Table II by the transpose of the inverse of T_2 (Table IV). This transformation may be termed T_3.

The difference between the factor loadings obtained by the two transformations can be expressed as follows. With six factors there are six hyperplanes each containing the axes of five factors and the axes themselves can be regarded as the six lines in which these hyperplanes intersect one another. Since the factors are all oblique, each axis is not orthogonal to the hyperplane which does not contain it, but forms an angle with the normal to that hyperplane. In Thurstone's procedure, where guidance is sought from the conception of simple structure, the hyperplanes are first determined, and from these the normals and primary factors are derived. The orthogonal projection of the variables gives the structure of the factors, and the parallel projection gives the pattern—in which the simple structure is supposed to be found. The three sets—the hyperplanes, the normals to the hyperplanes or reference vectors as Thurstone calls them, and the primary factors—are related in such a way that if any one set is determined, the others are also determined. Thurstone, as has been said above, determines the hyperplanes first. But it is possible to begin elsewhere, and we have begun with the primary factors. As we have seen, the post-multiplication of F_1 by T_2 gives the factorial structure, i.e., the loads obtained by orthogonal projection on to the primary axes (the correlations between the variables and the factors). The post-multiplication of F_1 and T_3 gives the parallel projection of the variables on to the primary axes. There is, however, an advantage in making this parallel projection not on to the primary axes themselves but on to the reference vectors. To do this the loadings given by $F_1 T_3$ would have to be post-multiplied by the cosines of the angles between the primary axes and their corresponding reference vectors.

Each of the foregoing factorial arrangements has its own value. If simple structure is sought, parallel projection, of course, is essential. But on the whole, when knowledge is incomplete and factors are only in the process of discovery, preference, we think, should be given to the orthogonal projection of the variables on to the primary axes, i.e., to the factor loadings described as F_2. Our reasons may be briefly summarized as follows:

(1) In parallel projection, although the loads on the axes can be accurately calculated if the whole system is known, the factors themselves, on the whole, are difficult to identify and recognize

in the concrete. As in the case of the later factors in an orthogonal system, each factor is, as it were, what is left when several other things have been taken away; and generally it is not easy to realize what this residuum is: whereas with immediate orthogonal projection on to the primary axes, the psychological meaning of the factors can usually be grasped more readily.

(2) The successful use of parallel projection requires that the factor analysis has been completed and all the factors found. If, for example, the centroid analysis has extracted only five factors, when six are really present, and we attempt to find a factor-pattern on five oblique factors, the whole scheme goes awry; for even if good guiding points for the five axes can be obtained, the direction of projection on to them is not known. The hyperplane, parallel to which we project, contains only $(n-2)$ instead of $(n-1)$ dimensions. Of course, with direct orthogonal projection on to the axes representing the primary factors, there is also error; but the error is usually not so great. In this case what is missing is the guidance which would have been obtained from an additional centroid factor. With parallel projection, however, there is no guarantee that the missing factor is smaller than the rest; and moreover, the axis of each of the others is slightly out of its true position.

(3) It may well be that we cannot place all the primary factors correctly, even if we can extract the right number by the centroid process. In such a case parallel projection is unworkable; whereas the method of orthogonal projection on to the primary axes, when these axes are positively determined, may be used for any axes that are known. If an adequate centroid analysis has been made, a study of a single factor can be carried out with considerable success, even if the other factors remain undetermined.

V. THE FACTORS

The six factors which we have found, together with their relations to the original variables, may be set forth as follows, the figures in brackets indicating the correlation between the factor and the relevant variable.

Factor I—Spontaneity. With a high degree of this quality the subject is full and free in his reactions; he is brimming over, full of life, and is spontaneous and responsive in most situations. On the other hand with a low degree of it he is restricted, inhibited or apathetic and unable to let himself go. The spontaneous person tends to be (*a*) cheerful (·708); (*b*) a good mixer (·665); (*c*) free from the tendency to withdraw (·631); (*d*) excitable (·609); (*e*) likes stimulating and exciting things (·601); (*f*) is not shy (·511); (*g*) is responsive to the bright side (·496); (*h*) expresses emotions readily (·457); (*i*) is easily stirred up (·435), and (*j*) tends to recover his cheerfulness easily (·400).

Factor II—Stability. With a high degree of this quality the subject is well integrated, balanced, and mature. The stable person is (*a*) not generally emotional (·627); (*b*) nor excitable (·576); (*c*) nor easily flurried (·566); (*d*) restrains expression of emotions (·558); (*e*) is usually self-reliant (·544); (*f*) reliable (·544) and (*g*) faces up to problems (·528), and to tests (·488); (*h*) requires a high standard of behaviour (·487), and (*i*) has fixed ways of looking at things (·476); (*j*) is slow to blame (·460); (*k*) free from fits of depression (·456); (*l*) persevering and persistent (·432), and (*m*) not irritable (·431).

Factor III—Persistence. With a high degree of this quality the subject tends to carry through tasks which he has begun. He is not easily put off by difficulties, nor does he abandon activities quickly and readily when the first novelty has worn off. The persistent person (*a*) tends to face up to the problems of life (·708), and to tests (·591); (*b*) is usually balanced (·632); (*c*) self-reliant (·628), and (*d*) reliable (·612); (*e*) not usually suggestible (·483), nor (*f*) conceited (·479); (*g*) slow to blame (·470), and (*h*) systematic (·468); (*i*) not easily rattled (·453), and (*j*) requires a high standard of behaviour (·445).

Factor IV—Assertiveness. With a high degree of this quality the subject is quick and ready to maintain his point of view. He is not inclined to let things pass with which he does not agree. The assertive person (*a*) is free from false modesty (·633), and (*b*) not commonly submissive (·627); (*c*) tends to dominate and lead (·609), and (*d*) is not shy (·565); (*e*) is indifferent to the views of others (·508); (*f*) is aggressive (·497), and (*g*) free from feelings of inferiority (·466); (*h*) is not suggestible (·461) but (*i*) contra-suggestible (·431); (*j*) ambitious (·456); (*k*) has his own way of looking at things (·425), and (*l*) faces up to the problems of life (·405).

Factor V—Sensitiveness. With a high degree of this quality the subject is easily touched and responds to the actions and emotions of other people. The contrast of tender versus tough indicates the two ends of the scale with fair adequacy. The sensitive person (*a*) is generally emotional and easily stirred up (·607); (*b*) has a high standard of behaviour for himself (·548), and (*c*) expects it of others (·413); (*d*) is distressed by the difficulties of others (·534); (*e*) is touchy (·479), and (*f*) is excitable (·407); (*g*) compliant and conciliatory (·319); (*h*) does not relax (·309); (*i*) is naturally kind (·308).

Factor VI—Inferiority. With a high degree of this quality the subject suffers from distressing feelings of inferiority. These may be covered and even compensated for by aggressive action; but they underlie much that he does. The person who has a sense of inferiority (*a*) has a low opinion of himself and his abilities (·596); (*b*) lacks cheerfulness (·540); (*c*) avoids the problems of life (·498); (*d*) is slow to assert himself (·470); (*e*) is prone to sad thoughts (·467); (*f*) touchy (·451), and (*g*) does not easily regain cheerfulness (·433); (*h*) is shy (·418); (*i*) often unbalanced and immature (·364), and (*j*) is a poor mixer (·360).

VI. THE INTERRELATIONS OF THE FACTORS

The interrelations of the factors have been given in Table V above. But since, to reach this result, the original data have undergone a number of computational processes, it is desirable to obtain some check of the reliability and objectivity of the analysis as a whole. To this end use has been made of fresh material. A considerable time after the original experiment was carried out, it was repeated with different observers, reporting on 62 new subjects. Then, each factor being taken in turn, weights were provisionally allotted to the qualities associated with them as set forth in section V above, and scores were calculated for each subject. These, when inter-correlated, gave the following results:

TABLE VI. INTERRELATIONS OF FACTORS

Trait	1	2	3	4	5	6
1. Spontaneity..	1·000					
2. Stability	−·181	1·000				
3. Persistence	−·087	·587	1·000			
4. Assertiveness	·117	·182	·220	1·000		
5. Sensitivity	·039	−·120	−·036	−·174	1·000	
6. Inferiority	−·490	−·240	−·264	−·295	·120	1·000

The agreement of this table with Table V is probably as good as can be expected with the data and the method adopted. All significant figures have the same signs; most of the numerical values agree fairly well, the differences reaching the 5 per cent. level of significance only in two cases. Taking the two tables together the following generalizations may be ventured.

Spontaneity tends to be associated with assertiveness and to reduce stability and inferiority. Stability tends to be associated with persistence and possibly with assertiveness, and to reduce spontaneity, sensitiveness, and inferiority. Assertiveness tends to be associated with spontaneity, persistence, and possibly with stability, and to reduce inferiority and sensitiveness. Sensitiveness is possibly associated with inferiority, and tends to reduce stability and assertiveness. Inferiority tends to reduce assertiveness, persistence, stability, and spontaneity.

When a system of oblique axes is used, it is desirable to ask whether the inter-correlations of the factors can be accounted for by a secondary factor or set of such factors. But in the present instance at least three secondary factors would be required; and we have so far discovered no convincing positions into which their axes could be rotated. All the many positions which they might take up are equally arbitrary; and, in the light of existing knowledge, it seems inadvisable to lay any weight on these factors.

The ultimate value of factors derived by analysis lies in the light which they throw on the individual. They provide a framework within which the study of him may be set, and obviously the first step will be to assign him a score in respect of each factor. This, however, does not exhaust the situation. There is what may be called a normal relation between the factors themselves, discovered in the first instance statistically by reference to the total population examined. But there is no ground for believing that this statistical or normal relation must hold in detail for every individual. The interrelations of the factors in each individual are also a part of the true picture of him, and those relations merit examination.

The most obvious method would be the direct one. The individual may be measured a considerable number of times in respect of each of the factors, and the correlations resulting from those measurements may be compared with the normal values: that is to say, something analogous to what Burt has called P-technique (1, pp. 169–209; 2, pp. 178–203) may be applied, not to discover factors, but to investigate the relations between them in the individual. The process would be laborious; and, although it seems the most fundamental method, it is not always practicable.

Without applying it generally, it may perhaps be possible by considering individual results to gain some idea where further analysis on these or other lines is desirable. When the score of an individual on a factor deviates significantly from the mean, we can ask whether this deviation is accompanied by any other unexpected deviation in his scores on other factors. In the case of orthogonal factors, the normal exception is the mean itself; and hence in this case the question which we have to ask is whether more factors than one deviate from the mean at the same time. If two or more such factors do so, this may be due to accidental variations in the separate factors; but it may also be due to an unusual connexion between the factors in the individual. In the case of oblique factors, when one deviates from the mean, the most probable position of the other is not the mean itself, but a distance from it given by a regression equation. If the other factors do not have the values which these considerations lead one to expect, a qualitative analysis of the individual is generally indicated. If the factors obtained by analysis are used in this way their value may be greatly enhanced. But a fuller discussion of the point must be reserved for another occasion.

VII. SUMMARY

1. The purpose of the experiment discussed in this paper was to modify and extend the analysis of personality qualities already made by the method used by the authors. Eighty observers each gave a personality sketch of two subjects, rating them also on 45 or 46 qualities. The resulting correlations yielded six centroid factors.

2. Rotation of the centroid axes was undertaken both into orthogonal and into oblique positions, the latter being dealt with in two ways, viz., to give structure on 'primary factors' and also on 'reference vectors.' The reasons for selecting the positions into which the axes were rotated have been indicated above. In the end preference was given to the method which projects orthogonally on to oblique primary axes.

3. It is claimed that the resulting factors and their inter-correlations may provide a framework within which the study of the individual may be set.

REFERENCES

1. Burt, Cyril (1940). *The Factors of the Mind.* University of London Press.
2. Burt, Cyril (1940). 'The factorial study of temperamental traits.' *Brit. J. Psychol., Stat. Sect.*, I, 178–203.
3. Cattell, R. B. (1933). 'Temperament tests.' *Brit. J. Psychol.*, XXIII, 308–29.
4. Cattell, R. B. (1944). 'Interpretation of the twelve primary personality factors.' *Character and Personality*, XIII, 55–91.
5. Cattell, R. B. (1946). *Description and Measurement of Personality.* Harrap and Co.
6. Cattell, R. B. (1947). 'Confirmation and clarification of primary factors.' *Psychometrika*, XII, 197–220.
7. Garnett, J. C. M. (1918). 'General ability, cleverness and purpose.' *Brit. J. Psychol.*, IX, 345.
8. Guilford, J. P., and Guilford, R. B. (1939). 'Personality factors D, R, T and A.' *J. Abnorm. Soc. Psychol.*, XXXIV, 21–36.
9. Reyburn, H. A., and Taylor, J. G. (1939). 'Some factors of personality.' *Brit. J. Psychol., Gen. Sec.*, XXX, 151–65.
10. Reyburn, H. A., and Taylor, J. G. (1941). 'Factors in introversion and extraversion.' *Ib.*, 147-9;
11. Reyburn, H. A., and Taylor, J. G. (1941). 'Factorial analysis and school subjects: a Criticism.' *Trans. Roy. Soc. S.A.*, XXVIII, 168–85.
12. Reyburn, H. A., and Taylor, J. G. (1943). 'On the interpretation of common factors: a criticism and a statement.' *Psychometrika*, VIII, 53–64.
13. Reyburn, H. A., and Taylor, J. G. (1943). 'Some factors of temperament: a re-examination.' *Psychometrika*, VIII, 91–104.
14. Reyburn, H. A., and Raath, M. J. (1949). 'Simple structure: a critical examination.' *Brit. J. Psychol., Stat. Sect.*, II, 125–33.
15. Thomson, Godfrey H. (1948). *The Factorial Analysis of Human Ability.* University of London Press.
16. Thurstone, L. L. (1945). *Multiple-Factor Analysis.* University of Chicago Press.
17. Webb, E. (1915). 'Character and intelligence.' *Brit. J. Psychol. Mon. Suppl.* No. 3.
18. Wolfle, D. (1942). 'Factor analysis in the study of personality.' *J. Abnorm. Soc. Psychol.*, XXXVII, 393–7.

THE SIGNIFICANCE OF FACTOR LOADINGS

LAWLEY'S TEST EXAMINED BY ARTIFICIAL SAMPLES

By STEN HENRYSSON

Department of Psychology and Education, University of Uppsala, Sweden

I. *Introduction.* II. *The Construction of Artificial Samples.* III. *Conclusion.*

I. INTRODUCTION

One of the weak points in factor analysis is the lack of satisfactory significance tests[1] for deciding whether the residuals left in the correlation matrix after extracting a certain number of factors can be explained by sampling errors or not. A step towards a solution of this problem was taken by Lawley (1940, 1941, 1943). Starting (1940, p. 65) from Wishart's joint probability distribution of the sample variances and covariances,[2] he finds a likelihood function, which is treated as a function of the population parameters (the factor loadings) and for this he endeavours to find a practicable solution.

His analysis starts with a preliminary estimate of the number, m, of common factors, and loadings for these factors are then found by an iterative process, beginning with a set of trial values. A method, suitable for large samples, is suggested for testing the assumption initially made in regard to the precise number of factors required. From his likelihood function he deduces that, if the assumption is correct, then the quantity

$$N \sum_{i<j} \left\{ \frac{(c_{ij} - a_{ij})^2}{s_i{}^2 s_j{}^2} \right\} \text{ , } (i, j = 1, 2, 3, \ldots, n) \qquad (1)$$

(where N is the number of observations, c_{ij} are the covariances calculated from the loadings, a_{ij} are the given covariances, and $s_i{}^2$ are the error variances) will be distributed as χ^2 with p degrees of freedom, where

$$p = \tfrac{1}{2}\{(n - m)^2 - (n - m)\} \text{ .} \qquad (2)$$

"If a significantly high value of χ^2 is obtained, this will indicate that we must reject the hypothesis and assume the existence of more than m factors" (1941, p. 184). This test of significance is applicable only if the above method of estimation is used.

The object of the following paper is to examine Lawley's test of significance by means of artificial samples. The test is intended to be used on large samples; and our object is to find out if it works on samples of 200 observations. This will be done by ascertaining how the values of χ^2, which we get for our different samples, are distributed.

II. THE CONSTRUCTION OF ARTIFICIAL SAMPLES

Wold (1948) gives 25,000 numbers taken at random from a normally distributed population ($M = 0$, $\sigma = 1$). The numbers are given in ten uncorrelated columns, $x_1, x_2, \ldots, x_{10}$. We construct nine variables containing one general factor according to the factor pattern

$$z_i = x_1 + x_{i+1} \quad (i = 1, 2, \ldots, 9) \qquad (3)$$

---

[1] See, however, the paper by Professor M. S. Bartlett in the last number of this *Journal* (III, 1950, pp. 77–85) on 'Tests of Significance in Factor Analysis,' especially sect. II.

[2] The method is not restricted to correlation coefficients.

by adding to the number in column (1) first, the number in column (2), secondly, the number in column (3), and so on until we reach column (10). The value in column (1) is thus common to all the nine variables. We divide the first 24,000 random numbers into twelve parts with 200 rows in each, and thus get *twelve* samples of nine variables each containing 200 observations :

$$z_{iN} = x_{1N} + x_{(i+1)N} \quad (N = 1, 2, \ldots, 200) \qquad (4)$$

In each sample of 200 observations the variances and covariances of the variates z_i are given by the formula

$$a_{ij} = \frac{1}{199}\left[\Sigma z_i z_j - \frac{1}{N}(\Sigma z_i)(\Sigma z_j)\right] (i, j = 1, 2, \ldots, 9), \qquad (5)$$

where Σ here denotes summation over the sample. The calculations may be simplified by using equation (3). Thus we have

$$a_{ij} = \frac{1}{199}\left[\Sigma x_1^2 + \Sigma x_1 x_i + \Sigma x_1 x_j + \Sigma x_i x_j - \frac{1}{200}(\Sigma x_1 + \Sigma x_i)(\Sigma x_1 + \Sigma x_j)\right]. \quad (6)$$

The sums in the equation above (which are sums over 200 numbers) may be obtained by adding together sums over four successive groups of 50 numbers. These latter have been calculated by Wold (1948) and Azorin and Wold (1950). For the first sample of 200 the matrix of variances and covariances [1] was found to be as follows :

TABLE I. MATRIX OF VARIANCES AND COVARIANCES

Variable	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)
(i)	3·894	1·819	2·192	1·957	1·878	1·987	1·845	2·091	1·959
(ii)	1·819	3·833	1·829	1·688	1·757	1·794	1·503	1·769	1·562
(iii)	2·192	1·829	4·136	2·291	1·779	2·066	1·793	2·406	1·896
(iv)	1·957	1·688	2·291	3·948	2·069	1·968	1·787	1·957	1·835
(v)	1·878	1·757	1·779	2·069	3·885	1·867	1·716	1·864	2·012
(vi)	1·987	1·794	2·066	1·968	1·867	3·739	1·917	2·003	1·815
(vii)	1·845	1·503	1·793	1·787	1·716	1·917	3·358	1·608	1·835
(viii)	2·091	1·769	2·406	1·957	1·864	2·003	1·608	3·993	1·907
(ix)	1·959	1·562	1·896	1·835	2·012	1·815	1·835	1·907	3·417
Loadings	1·438	1·230	1·495	1·419	1·354	1·402	1·264	1·428	1·343

Assuming one general factor, the loadings estimated by the maximum likelihood method were as shown in the last line of the table. The value of χ^2 required to test the hypothesis is found to be 21·8, which, with 27 degrees of freedom, is not significant. The probability P that such a value would be exceeded by chance is 0·74. For the whole set of twelve samples the values of χ^2 and the corresponding values of P are as given below.

TABLE II. VALUES OBTAINED FOR CHI-SQUARED

Sample	1	2	3	4	5	6	7	8	9	10	11	12
χ^2	21·8	38·6	24·2	19·2	28·1	17·8	31·6	31·9	28·2	22·1	25·8	15·5
P	0·74	0·07	0·62	0·86	0·41	0·92	0·25	0·24	0·40	0·73	0·53	0·96

---

[1] When calculating the variances and covariances by equation (6), we divide by 100 instead of 199 to make this step easier. This operation has no influence on the final values of χ^2.

III. CONCLUSION

The results obtained are in good agreement with theory. The values of P should conform with a rectangular distribution ; and the expected range is 0·85 (Cramér, 1946, p. 372), a figure which is close to the observed range (0·96 — 0·07) of 0·89. If all twelve samples are combined, a total value for χ^2 of 304·8 is reached. This, with $12 \times 27 = 324$ degrees of freedom, corresponds to a probability P of 0·77. From the test we should therefore conclude that all the 9×9 covariance matrices can be expressed in terms of one significant common factor only, with nine specific factors : and this conclusion accords with what in these artificial cases we know to be the fact.[1]

REFERENCES

1. Azorin, F., and Wold, H. (1950). 'Product sums and modulus sums of H. Wold's normal deviates.' *Trabajos de Estadistica*, I, 5–29.
2. Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton : Princeton University Press.
3. Lawley, D. N. (1940). 'The estimation of factor loadings by the method of maximum likelihood.' *Proc. Roy. Soc. Edin.*, LX, 64–82.
4. Lawley, D. N. (1941). 'Further investigations in factor estimation.' *Proc. Roy. Soc. Edin.*, LXI, 176–85.
5. Lawley, D. N. (1943). 'The application of the maximum likelihood method to factor analysis,' *Brit. J. Psychol.*, XXXIII, 172–5.
6. Thomson, G. H. (1948). *The Factorial Analysis of Human Ability*, Ch. xxi. London : University of London Press.
7. Wold, H. (1948). 'Random normal deviates,' *Tracts for Computers*, XXV.

APPENDIX

By CYRIL BURT

Maximum Likelihood Loadings as Obtained from Single-Factor Correlation Tables

Mr. Sten Henrysson, in his suggestive paper, has applied Lawley's test of significance to a set of factor loadings calculated by his first 'method of maximum likelihood' from an artificial covariance matrix containing one common factor only. In this country factor analysis has been most frequently carried out, not with covariances, but with correlations. Accordingly, as many readers of this *Journal* may be interested in that form of application, I have ventured to convert Mr. Henrysson's matrix of covariances [2] into a matrix of correlations. The coefficients so obtained are set out in Table III, with the corresponding factor-saturations added at the foot.

A word or two is perhaps required to indicate how British psychologists have commonly approached such a problem. In early discussions on the statistical validation of mental tests, interest centred almost entirely on the 'general ability' (or 'intelligence'), which seemed to dominate all others. Accordingly in 1915 I ventured to put forward a simplified procedure for analysing correlation tables which contained no more than a single significant factor, i.e., tables of the type which Mr. Henrysson has here constructed. In such cases it seemed clear that, for the purposes of the psychologist, Pearson's method of obtaining 'the line of closest fit' by the method of least squares needed modification. When we are dealing not with physical but with psychological measurements (the results of different intelligence tests, for example), we are almost bound to assume that a unique or specific factor is present in each test. The larger the specific factor, the more it tends to reduce what (in a somewhat broad sense) was originally called the 'reliability' or 'precision' of the test. From this it was argued that, in accordance with the usual least squares procedure, the correlation table should in theory first be 'prepared' by dividing each standardized variable throughout by the square root of its unreliability. And the argument was later extended to tables in which several common factors might be involved. In the case of an *ideal* single-factor matrix it was easy to show that this theoretical correction led to the same figures for the factor-saturations as would be obtained by applying Pearson's procedure direct to the correlation table with *reduced self-correlations* in the principal diagonal instead of unity; and further that, in such a case, this procedure itself could be further simplified by substituting the formula for simple summation for the more elaborate calculations required by Pearson's method of principal axes. It was therefore suggested that, even in dealing with an *empirical* table where the observed coefficients did not differ appreciably from a perfect hierarchical arrangement, the same shortened procedure might legitimately be adopted.[1]

---

[1] I should like to thank Professor Wold for his help and encouragement in this research, which was carried out in the Department of Statistics in the University of Uppsala, and Dr. D. N. Lawley and the editors, Professor Thomson and Professor Burt, for their assistance in preparing my manuscript for publication.

[2] Mr. Henrysson states (footnote p. 160) "when calculating the variances and covariances, we divide by 100 instead of by 199" (or 200). Hence to obtain variances and covariances as calculated by the more usual procedure, his figures would have to be halved or (more strictly) to be multiplied by 100/199.

As applied to the problem of fitting a table of correlations the method of simple summation was identical with that already in use for fitting a contingency table (1, p. 64, eq. 1); and it was natural to suggest that the discrepancies or 'residuals' should be tested for significance in the same familiar fashion, namely, by the chi-squared test (1, p. 64, eq. 3). On the same grounds as before, however, it appeared desirable to retain the correction for specificity for the residuals before testing their significance in this way. This resulted in a formula for residual correlations which was identical in form with the Pearson-Yule formula for partial correlation: (cf. 2, p. 57). The similarity, of course, is not accidental: precisely the same principles are involved in the two arguments.

The chi-squared test furnishes a significance test for the system of residuals taken as a whole, and was thus used chiefly with a bipolar analysis. To locate significant residuals occurring either in isolation or in clusters, it is more instructive to test each separately by an appropriate probable or standard error; this method was accordingly used in group factor analyses. Since each augmented residual (or 'specific correlation,' as it was called) could be regarded as a partial correlation with the general factor eliminated, it seemed legitimate (with large samples) to test it by taking the ordinary standard error for a partial correlation of zero, that is, by taking $1/\sqrt{N}$, where N denotes the number of persons tested: (1, p. 352, eq. 15). This implies that the augmented residuals can be treated as conforming approximately to a normal distribution with a standard deviation of $1/\sqrt{N}$; and, on that assumption, the sum of their squares, multiplied by N, would conform to the chi-squared distribution. Thus, without any very rigorous proof, it was assumed, and the assumption seemed borne out in practice, that the two procedures would yield much the same results.

The object of this note is to show that, in cases like that selected by Mr. Henrysson, the results obtained by this simple procedure[2] are identical with those obtained by the more elaborate method of calculation which he has adopted in his paper.

---

[1] This procedure was first briefly outlined in a paper, submitted to the Psychological Subsection of the British Association, on 'General and Specific Factors Underlying the Primary Emotions': (the summary, as printed in *Brit. Ass. Ann. Rep.*, 1915, pp. 694–6, gives the figures so obtained, and discusses the concrete results, but does not describe the technicalities of the method in any detail; a fuller account was given in my L.C.C. Report on 'The Relations between Ability in Different Subjects in the School Curriculum,' reprinted, with some abridgment, in *The Distribution and Relations of Educational Abilities*, 1917).

It is perhaps necessary to remind those who refer to these earlier papers that the use of terms has since slightly changed. In conformity with Galton's terminology, the phrases 'specific abilities,' 'specific factors,' and 'specific correlations' were originally employed to refer to what would now be called 'group factors.' The phrase 'specific factor,' however, was later adopted by Spearman and his pupils to denote a factor peculiar to a single test, i.e., what is now sometimes called a 'unique factor'; and 'reliability' was used to mean, not 'precision,' but 'self-consistency.' Hence what was originally called 'correction for unreliability' was later called 'correction for specificity.'

[2] I am tempted to suggest that the short and simple algebraic proof, put forward to support the older method, may perhaps be easier for the student to follow than the more rigorous approach adopted by Lawley. The assumptions made in the text imply that the matrix to be analysed is not (in the usual notation) R, the matrix of observed correlations with unity in the diagonal, but $R_u^{-\frac{1}{2}}(R-R_u)R_u^{-\frac{1}{2}} = R_u^{-\frac{1}{2}}R_cR_u^{-\frac{1}{2}} = R_a$ say, where R_u denotes the diagonal matrix of specific factor variances, and $R_c = R - R_u$, the correlation matrix with reduced self-correlations. This formulation leads to equations of the usual type, namely,

$$f_a'(R_a - v_aI) = 0 \text{ and } |\,R_a - v_aI\,| = 0\,; \qquad \textbf{(i)}$$

but the formula now relates to correlations corrected for 'unreliability.' And the procedure requires, not Spearman's correction formula $\frac{r_{ij}}{\sqrt{r_{ii}\,r_{jj}}}$, but the complementary formula, $\frac{r_{ij}}{\sqrt{1-r_{ii}}\,\sqrt{1-r_{jj}}}$, where r_{ii} and r_{jj} denote 'reliability' interpreted as above (viz., 'reliability' so far as the tests i and j depend on common factors only).

TABLE III. THE 'OBSERVED' CORRELATIONS

Variable	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)
(i)	(·5314)	·4708	·5462	·4991	·4828	·5207	·5102	·5303	·5370
(ii)	·4708	(·3950)	·4594	·4339	·4553	·4739	·4189	·4522	·4316
(iii)	·5462	·4594	(·5392)	·5670	·4438	·5254	·4811	·5920	·5043
(iv)	·4991	·4339	·5670	(·5102)	·5283	·5122	·4908	·4929	·4996
(v)	·4828	·4553	·4438	·5283	(·4736)	·4899	·4751	·4733	·5522
(vi)	·5207	·4739	·5254	·5122	·4899	(·5277)	·5410	·5184	·5078
(vii)	·5102	·4189	·4811	·4908	·4751	·5410	(·4749)	·4391	·5417
(viii)	·5303	·4522	·5920	·4929	·4733	·5184	·4391	(·5085)	·5163
(ix)	·5370	·4316	·5043	·4996	·5522	·5078	·5417	·5163	(·5288)
Simple Summation	·7302	·6285	·7335	·7140	·6887	·7268	·6886	·7123	·7273
Weighted Summation	·7290	·6285	·7343	·7143	·6882	·7264	·6891	·7131	·7272
Loadings:									
from Correlations	1·438	1·230	1·493	1·419	1·356	1·404	1·263	1·425	1·344
from Covariances	1·438	1·230	1·495	1·419	1·354	1·402	1·264	1·428	1·343

In the present case the working is as follows. Starting with the ordinary table of correlations (Table III), saturations have first been computed by 'simple summation' (last line but three in the table). These have then been taken as a basis for obtaining the improved values furnished by 'weighted summation' (last line but two). It may be noted that, although the matrix is not perfectly hierarchical, the use of weighting introduces hardly any change in the figures. Each saturation has then been multiplied by the square root of the variance as given in Mr. Henrysson's table. Thus for variable (i) we obtain $\cdot 7290 \times \sqrt{3\cdot 894} = 1\cdot 438$, and so on. It will be seen that (except for trifling discrepancies) *the loadings thus finally reached are identical with those obtained by Mr. Henrysson from his matrix of variances and covariances* (see last two lines of Table III).

Next, taking the improved values for the saturations, a 'hierarchical' matrix of correlations has been constructed in the usual way to fit the observed correlations. The differences

Now, if R_c is a matrix of rank one, the ordinary method of expanding the characteristic determinant leads at once to

$$0 = | R_c - v_c I | = v_c^n - (r_{11} + r_{22} + \ . \ . \ . + r_{nn}) \, v_c^{n-1} ,$$

which gives $v_c = 0$ or $v_c = \Sigma r_{ii} = \Sigma r_{ig}^2 = f_c' f_c$; and similarly $| R_a - v_a I |$ leads to $v_a = 0$ or $v_a =$ $f_a' f_a$. But $f_a f_a' = R_a = R_u^{-\frac{1}{2}} R_c R_u^{-\frac{1}{2}} = R_u^{-\frac{1}{2}} f_c f_c' R_u^{-\frac{1}{2}}$. Hence we may take $f_a' = f_c' R_u^{-\frac{1}{2}}$ and $v_a = f_c' R_u^{-1} f_c$. From this it is easy to show that, under these conditions, the results of factorizing R_c will satisfy equations (i). Substituting the values thus found for f_a' and v_a, we have

$$f_a' R_a = f_c' R_u^{-\frac{1}{2}} \, . \, R_u^{-\frac{1}{2}} R_c R_u^{-\frac{1}{2}} = f_c' R_u^{-1} f_c \, . \, f_c' R_u^{-\frac{1}{2}}$$
$$= f_c' R_u^{-1} f_c \, . \, f_a' = v_a f_a' :$$

that is, $f_a' (R_a - v_a I) = 0$, as required. Thus, when the reduced correlation matrix is hierarchical, there is no need to carry out the elaborate calculations entailed by equations (i): the observed correlation matrix can be factorized just as it stands, except that reduced self-correlations must first be substituted for unity. Since in such a case $r_{ii} = r_{ig}^2$, the expression for the residuals becomes $\dfrac{r_{ij} - r_{ig} r_{jg}}{\sqrt{1 - r_{ig}^2} \sqrt{1 - r_{jg}^2}}$. (Note that, if R_c involves *several* common factors, $F_c' R_u^{-1} F_c$ is *not* a diagonal matrix, and thus does not satisfy the 'orthogonality condition' in the form imposed by Lawley.)

between these theoretical figures and the observed figures in Table III (i.e., the first ' residuals ') have then been divided by the product of the square roots of the two relevant specific factor variances. Thus

$$r_{12\cdot g} = \frac{\cdot 4708 - \cdot 7290 \times \cdot 6285}{\sqrt{1 - \cdot 7290^2}\,\sqrt{1 - \cdot 6285^2}} = \cdot 0238.$$

These ' corrected residuals ' are shown in Table IV.

TABLE IV. THE ' SPECIFIC ' CORRELATIONS $\left(r_{ij\cdot g} = \frac{r_{ij} - r_{ig}r_{jg}}{\sqrt{1 - r_{ig}^2}\,\sqrt{1 - r_{jg}^2}}\right)$

Variable	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)
(i)	—	·0238	·0232	−·0449	−·0376	−·0183	−·0157	·0215	·0149
(ii)	·0238	—	−·0042	−·0276	·0405	·0327	−·0254	·0070	−·0476
(iii)	·0232	−·0042	—	·0892	·1249	−·0172	−·0510	·1430	·0637
(iv)	−·0449	−·0276	·0892	—	·0727	−·0135	·0028	−·0341	·0431
(v)	−·0376	·0405	·1249	·0727	—	−·0194	·0017	−·0342	·1044
(vi)	−·0183	·0327	−·0172	−·0135	−·0194	—	·0812	·0009	−·0428
(vii)	−·0157	−·0254	−·0510	·0028	·0017	·0812	—	−·1035	·0816
(viii)	·0215	·0070	·1430	−·0341	−·0342	·0009	−·1035	—	−·0054
(ix)	·0149	−·0476	·0637	·0431	·1044	−·0428	·0816	−·0054	—

With 200 cases the standard error for a partial correlation of zero may be taken as $1/\sqrt{200} = 0{\cdot}0707$. Of the 36 figures in Table IV, only one is more than twice the standard error, which is in agreement with Mr. Henrysson's value for $P\,(0{\cdot}74)$. Chi-squared can be calculated by squaring each of the ' corrected residuals,' adding the squares, and multiplying by $N = 200$. The figure so obtained is 22·0 as compared with Mr. Henrysson's figure of 21·8. Thus chi-squared for the residual correlations is practically identical with chi-squared for the residual covariances, as was after all to be expected. (Here, and in the foregoing figures for the loadings, the slight discrepancies between his figures and my own are no larger than might be anticipated from a difference in the extent to which the iterative calculations have been carried. Indeed, had we dispensed with iterations altogether and merely taken the figures given by simple summation, the result would have been virtually the same.)

Tables III and IV above correspond respectively with Tables XVIII and XX and Tables XXII and XXIII of my early Report (2, 1917, pp. 52 and 57 and pp. 61 and 62). The factor-saturations there given were derived by inserting reduced self-correlations in place of the ordinary reliability coefficients or self-correlations of unity.

The use of reduced self-correlations was strongly criticized at the time by both Pearson and Spearman, and has since again been questioned in the last issue of this *Journal* by Professor Thomson.[1] But it may, I think, be claimed that the agreement between the results reached by this substitution and those furnished by the first ' method of maximum likelihood ' can be plausibly adduced as one further argument in favour of reduced self-correlations, at any rate in such cases as the present.

The variations in the saturation coefficients are themselves not without interest. It will be remembered that, in constructing his covariance table, Mr. Henrysson gave an equal weight to the general and to the specific factor in every case. Hence the expected value for each of the nine saturations will be $1/\sqrt{2} = 0{\cdot}707$. The standard deviation of the observed values about this true value is 0·032. With 200 cases the standard error of an ordinary correlation of 0·707 would be ±0·035. At first sight the similarity between the two figures might seem to support the suggestion that a rough estimate of the standard error for the saturation coefficient can be obtained by regarding it as an ordinary coefficient of correlation. However, figures obtained from dice-throwing experiments by Mr. H. F. Weston and myself some time ago indicate that such an estimate is decidedly too low. The approximate formula I myself have suggested (*Mental and Scholastic Tests*, p. 139) here gives a value of ±0·058. But further investigations along these lines would be highly instructive.[1]

---

[1] In *Human Ability* (1950) Spearman repeats his criticism of " the filling of the vacant diagonal places by stopgaps," which he thinks " quite arbitrary " (pp. 29–30). For Thomson's criticisms see *Brit. J. Psychol., Stat. Sect.*, III, 1950, p. 126, and refs.

REFERENCES

1. Yule, U. (1912). *Introduction to the Theory of Statistics*. London : Griffin.
2. Burt, C. (1917). *Report on the Distribution and Relations of Educational Abilities*. London : P. S. King.

---

[1] In a study of significance tests Mr. Weston has calculated artificial tables using *different* proportions for the general factor in each test, and including in some cases a second bipolar factor. Here too the figures obtained by applying the least squares procedure (' weighted summation ') to the correlation table, with reduced self-correlations in the diagonal, prove to be very similar to the figures reached by the more laborious procedure described by Lawley and Emmett (this *Journal*, II, pp. 90–7 ; cf. Burt, *ibid.*, pp. 106, 113). With the chi-squared tests he finds the 'correction for specificity' rather more generous than the z-transformation. It would be instructive if Mr. Henrysson could give detailed figures for the loadings in *all* his twelve tables. At the same time it should be remembered (as Dr. Lawley points out in a note he has been good enough to send me) that the errors of estimation in the nine loadings would not be uncorrelated. He adds that according to his own formula the standard error should be about ±0·063.

THE FACTORIAL ANALYSIS OF QUALITATIVE DATA

By CYRIL BURT

Psychological Department, University College, London

I. *The Importance of Qualitative Data in Psychology.* II. *Alternative Statistical Techniques.* III. *The Treatment of Multiple Determinates.* IV. *A Factorial Analysis of Physical Attributes.* V. *Summary and Conclusions.*

I. THE IMPORTANCE OF QUALITATIVE DATA IN PSYCHOLOGY

The Form of the Data. In many investigations within the field of individual differences, the available data are expressed, not as quantitative measurements stating magnitude or degree, but in terms of classes or attributes which are essentially qualitative. Statistical psychologists, and particularly those who have worked with standardized tests and employed factorial methods, are often accused of ignoring the qualitative aspects of their problems. As a rule, their critics seem to assume that, because such observations are not recorded in the form of quantitative assessments, they are no longer amenable to quantitative treatment, and can therefore have no use or interest for the statistical investigator. That, however, is a patent fallacy. If they are to be tabulated, such observations, it is true, must be set down in what the schoolboy calls the 'nought-and-one' style, where each 'one' will signify that yet another individual belongs to the class named or possesses the attribute specified, and 'nought' will signify that he does not. But, when that has been done, it is a simple matter to count the ones and compare their sum with that of the ones and noughts added together: in this way we can readily summarize our observations in the form of a table of frequencies or probabilities; and these can manifestly be subjected to statistical treatment: (cf. 3, 4, 5).

In psychology the oldest and most familiar instance of this procedure is furnished by the marks given for the Binet-Simon tests. Here each examinee is in effect awarded a measurement of 'one' for every test he passes and 'nought' for every failure. A similar device has long been adopted by teachers in marking the simpler type of examination paper: each child's total score is obtained by counting every correct answer as 'one,' and adding up the total. So-called personality-tests—the Rorschach, the personal inventory, the biographical questionnaire, and enquiries about interests and attitudes, for example—are frequently scored in this way.

As actually observed, no doubt, most mental traits (probably as a result of their highly complex origin) take the form of continuously graded variables. But the patterns, the combinations, and the interactions between such traits (or between the statistical factors to which it is convenient to reduce them) often have a qualitative character, which cannot itself be expressed as a difference of degree. Moreover, individual psychology is not concerned merely with the description of individuals considered in isolation. It has to take into account the causal influences which affect each individual and his observable behaviour; and many of these causal influences are not themselves reducible to quantitative variations. For example, in my early studies of delinquency, it was often essential to take up points which lead to a classification rather than a measurement. Is the child illegitimate? Is he an only child? Is he living with his parents? Has he suffered from this or that disease? What is the lad's present occupation? These and many other questions cannot be answered by a figure on a rating scale, but only by a categorical 'yes' or 'no,' or (where there are several alternatives) by a 'this and therefore not that or any other.'

Nor are qualitative data of this nature confined solely to psychology. They occur in many related fields. In physical anthropology, for example, the investigator has constantly to deal with what statistical writers have termed 'attributes' as contrasted with 'variables.' Pearson's well-known paper on 'the correlation of characters not quantitatively measurable' (3) was prompted by the need to deal with qualitative traits like hair-colour and eye-colour. And, as a glance at the earlier chapters of Yule's text-book (6) will show, some of the stock illustrations of such procedures are still drawn from the anthropological field. More recently in biological research, and particularly in work on experimental genetics, the investigator has been concerned quite as often with the assignment of individuals to distinct classes as with the measurement of variable traits.

The distinction between 'measurement-data' and 'frequency-data' and between 'continuous distributions' and 'discrete distributions' has long been recognized in statistics. And the growing realization that mathematics is not restricted to the study of quantity has made it increasingly possible to discuss qualitative characteristics in the same rigorous way, and often along the same lines, as quantitative. On the wider issues I may perhaps refer the more advanced student to the contributions of French statisticians, who have dealt most fully with the problem.[1]

Now we must, I think, agree with our critics that, in studying individual differences whether in physical or in mental characteristics, factorial investigations have hitherto concentrated chiefly on quantitative measurements. It would, however, be plainly advantageous if all kinds of characteristics, qualitative as well as quantitative, could be included in such studies. And the chief purpose of this paper will be to indicate how factorial methods may still be applied, even when much of the data is expressed in a qualitative form.

The Nature of the Factors. As with quantitative variables, so with qualitative attributes the factors thus derived will be essentially principles of classification. Indeed, the factor analysis of attributes is related, much more obviously than the factor analysis of variables, to the modes of classification employed in the simpler sciences.

In botany, for example, the trained observer inspects in order the parts of the plant he is seeking to identify—its root, its stem, its leaf, its flower, and its fruit; and knows that each of these parts may exhibit one of a number of alternative forms or qualities. The veining of the leaf may be parallel or netted; the colour of the flower may be white, red, yellow, blue; and so on. Even where some of the features have numbers attached, e.g., 3, 4, or 5 petals, these numbers still specify, not differences in degree, but differences in kind. The diagnostic combinations of such attributes correspond to the factors.

The chief divergence between problems of classification in the simpler sciences and in the anthropological sciences is that in the former a far higher degree of certainty attaches to the inferences to be drawn. Thus *all* tulips have parallel veins; nearly all hawthorns have flowers that are white or red; and, with few exceptions, all clover-leaves are three-lobed. In chemistry exceptions are rarer still: all metals of the alkali class are univalent; all occur in the first group of the periodic table; all act as powerful bases; all are strongly electro-positive, and the strength of this tendency almost invariably increases with increasing atomic weight (here, we may note in passing, the classification turns in part on differences in quantitative variables as well as on qualitative attributes). On the other hand, we do not find that *all* children who fail in the Binet tests for Age VIII also fail in those for Age IX; nor do we find that all leptomorphic individuals are schizothymic or that pyknic individuals are nearly always cyclothymic. A correlation technique therefore becomes essential.

With qualitative variables as with quantitative, the most useful factors will be those that lead to a hierarchical classification; and the natural procedure will be to seek, in any particular 'domain' or 'universe of discourse,' first the factor which will account for the largest amount of individual variation, and then the factor which will account for the largest amount of residual variation, and so on. Thus, to adopt an expressive term familiar to those brought up on Aristotelian logic, a factor may be thought as a διαφορὰ εἰδοποιός [2]—a hypothetical 'differentiating principle,' constituting a species.' Remembering that 'species' is a relative term, we shall therefore expect the widest factor to represent the genus under investigation ; and the remaining factors to represent the differences between species, sub-species, and so on, in hierarchical order. However, as modern logicians have so often pointed out, the traditional notion of classification found in the older text-books of logic assumed a neatly arranged system of natural kinds—each genus sharply distinguished from all other genera, each species or type unequivocally separated from the rest ; whereas in actual fact there is a vast and often irregular amount of overlapping, and this overlapping is greatest in the more complex sciences. Hence, in my view, the only safe and speedy way to reach a satisfactory solution in such cases is to invoke the aid of a suitable statistical technique. And, if the statistician has not as yet developed an appropriate method which he can offer us, then the psychologist, however imperfectly, must set to work to devise his own.[1]

---

[1] E.g., the contributions by different writers in Borel, E. (ed.) *Traité du Calcul des Probabilités* (1924 onwards). Kelley remarks on the fact that, whereas earlier writers who touched upon the statistical treatment of psychological problems classified 'statistical series' into 'qualitative' and 'quantitative,' more recent writers (like Garrett, Thurstone, Engelhart, McNemar) have preferred to classify series into 'discrete' and 'continuous,' and that, as they use the terms, the distinctions are by no means identical. Quantitative scales may also be discrete; and, so far as such writers touch upon discrete distributions, they restrict themselves almost entirely to quantitative instances (cf. 13, pp. 62f.).

To avoid misunderstanding, let me emphasize at once that I do not for one moment believe that all the psychologist's observations can be adequately dealt with by just tabulating frequencies. In my view *the method of trait-assessment, whether qualitative or quantitative, should always be supplemented by the method of case-study*. Neither is complete without the other. The ancient controversy between 'clinical' and 'metrical' psychologists has arisen largely because each seems to have assumed that his own method was the sole method. How the two kinds of approach may be combined in a single research I endeavoured to show in my book on *The Young Delinquent*. And I should hold that almost every problem of individual psychology demands this double mode of treatment.

[2] Arist., *Top.* VI, vi, 2; cf. *Nicom. Eth.*, X, iv, 3.

II. ALTERNATIVE STATISTICAL TECHNIQUES

Correlation in the Case of Two Dichotomous Classifications. In my first attempt at factorizing results obtained with the Binet scale, each test was paired with every other, and the number of children passing both was taken as indicating the association between the abilities so tested. In view of the large number of tests in the scale, the frequencies for the several pairs were entered in a single table of double entry, like an ordinary contingency-table : the whole table was then fitted with theoretical values by means of the simple summation formula ; and the sums were taken as indicating the degree to which each test was representative of the common factor.[2] In later researches, when the factor analysis of correlation coefficients had become a recognized procedure, attempts were made to reduce the initial data to correlational form. Since cognitive abilities are continuous variables, the ideal procedure, so it was thought, was to modify the Binet problems, wherever possible, in order to secure graded measurements, and, where this was not possible, to calculate tetrachoric coefficients, on the ground that the distribution of such abilities conforms pretty closely with the normal curve.[1]

---

[1] The logical difficulties surrounding attempts at classification in the more complex sciences are perhaps most clearly brought out by Sigwart's discussion ; and his solution is in keeping with the procedure suggested above. By what principle, he asks, are we to be guided in our " progressive efforts to work by ascending generalization from the lower to the higher classes, or downwards from the higher to the lower " ? At each stage, he replies, " the most serviceable arrangement will be that which groups together those things which have the largest number of characteristics in common " (*Logic*, Part III, 'Technical' ; esp. pp. 158f. : it is suggestive to note that he proceeds to cite, as the cases showing the most difficult and complicated structure, the attributes or 'elementary characteristics' studied in psychology).

But how (he goes on to ask) are we to compile our preliminary list of characteristics in order to see what is the largest number that certain groups have in common ? That, as he points out, is plainly not a formal but a material problem. To which I need only add that, if it is a material problem, then it is not a problem that can be decided by statistical calculation alone.

There is, I am well aware, a widespread notion that, in dealing with quantitative measurements, the investigator can proceed by first taking all the tests that are applicable or throwing together all data that may be available, trusting that the statistical analysis will then " discover " the underlying factors. But that, I venture to assert, is quite fallacious. The initial collection of data is or should be guided by some prior hypothesis which the factor analysis is intended to check ; and the very process of collecting the data must imply some tentative ideas, conscious or unconscious, as to the nature of the classification which the investigator is testing. Most probably these ideas have resulted from some causal theory that has led him to suspect that these or those distinctive manifestations may be grouped together to form this or that species or genus. For example, in psychology neurological knowledge may have led him to anticipate a general cognitive factor ; or anthropological knowledge to expect a conjunction of certain shades of hair, eye, and complexion ; or again his psychiatric theories or his case-studies may have induced him to believe that certain signs or symptoms will cluster together. But whatever they may be, all such prior considerations will, as a rule, be based not on a statistical analysis, but on concrete observations or on speculative hypotheses.

[2] Cf. Yule, U., *loc. cit.*, p. 64, eq. 1, and p. 66, Table III. It was in fact this formula which suggested the use of 'simple summation' and the application of the chi-squared test of significance for correlation tables calculated from quantitative variables in the ordinary way.

However, as I have already pointed out, with much of the data obtained in work on non-cognitive differences, the information may be essentially qualitative in nature and available only in classificatory shape. In such cases the tetrachoric coefficient is quite unsuitable. Nevertheless, without making any specific assumptions about the nature or distribution of the underlying variables, it is still possible to express the association between such characteristics by means of correlation. Thus, to obtain what in form at least is a product-moment coefficient, Yule [2] suggested taking the line of division between each pair of classes as the arbitrary origin and the class interval as unit, and then calculating the mean, the standard deviations, and the product-sum in the usual way. Using the notation I have adopted in previous articles,[3] suppose that, in a population of $N = N_0 + N_1$, a given attribute is present in N_1 cases and absent in N_0 cases; then the mean will be $\frac{1}{2N}(N_1 - N_0) = \frac{N_0}{N} - \frac{1}{2}$, and the standard deviation $\frac{\sqrt{N_0 N_1}}{N}$. This is tantamount to assigning a value of $+\frac{N_0}{N}$ (the mean of the positive subgroup) to denote the *presence* of the attribute, and $-\frac{N_1}{N}$ (the mean of the other subgroup) to denote its *absence*. We can now compute the frequencies for joint deviations; and, with this or some equivalent device, we can always calculate a complete set of correlation coefficients between all possible pairs from a series of attributes. Then, if we wish, the table of coefficients can be factorized, and the attributes expressed in terms of the factors so obtained. This method—namely, the calculation of coefficients from all the 2 × 2 tables formed from the initial tables of frequencies—was adopted for a factorial study of the Binet tests, and appeared to give quite satisfactory results.[4]

Available Procedures. Evidently, then, there is no intrinsic difficulty about factorizing data of this kind. According to the problem at issue or the assumptions we prefer to make, various working methods may be adopted. From the standpoint of the research-student, perhaps the simplest way to classify the possible procedures is to begin with the familiar distinction between (A) methods appropriate to cases where *measurement* is at least conceivable and (B) methods appropriate to cases where it is more natural to think in terms of *frequencies*.

(A) Methods of the first type proceed by converting the initial data, explicitly or implicitly, into quantitative form, and then base the factor analysis on a table of correlations or of some equivalent type of coefficient. For the present purpose the most useful coefficients are the following:

- (i) the product-moment correlation for a twofold point distribution (φ);
- (ii) Yule's colligation coefficient (φ for an equalized table);
- (iii) Yule's association coefficient;
- (iv) Pearson's coefficient of contingency, corrected if necessary for degrees of freedom by Tschuprow's formula;
- (v) Pearson's tetrachoric coefficient;
- (vi) the correlation ratio.

---

[1] Cf. John, Enid, and Burt, C., 'Factor Analysis of the Terman Binet Scale,' *Brit. J. Educ. Psych.*, XII, 1942, p. 122. On the relative merits of different methods of calculating association coefficients for such data as are furnished by the Binet tests, cf. *Mental and Scholastic Tests*, pp. 137 and 232, also *The Backward Child*, Appendix III. As to the reasons for preferring graded measurements, see *id.*, 'The Measurement of Intelligence by the Binet Tests,' *Eugenics Review*, IV, 1914, pp. 140f.

[2] Yule, U., *Introduction to the Theory of Statistics*, 1912, pp. 216–7.

[3] E.g., Burt, C., 'Statistical Problems in the Evaluation of Army Tests,' *Psychometrika*, IX, 1944, p. 225, and refs.

[4] Cf. *Mental and Scholastic Tests*, p. 137. As I noted at the time, however, the use of an association coefficient or point-distribution correlation commonly introduces "an additional factor which seems to indicate the relative difficulty of the various tests' (*loc. cit.*, footnote 1). This factor, as a rule, does not appear when tetrachoric or ordinary product-moment coefficients are used, because they eliminate differences in the point of division.

(B) Methods of the second type leave the data in qualitative form. In dealing with such material we should, I think, frankly recognize that factor analysis may be legitimately applied to any kind of symmetrical table, not merely to correlations or covariances. In this case the figures to be factorized will mostly take one or other of the following forms :

(i) the absolute frequencies ;
(ii) the relative frequencies or probabilities ;
(iii) some other ratio, e.g., the square contingency or its derivatives.

To a large extent the distinctions on which the above classification rests are distinctions of procedure rather than of principle. Both the contingency coefficient and the root-mean-square contingency (φ) are closely related to the square contingency (χ^2) and to the frequencies on which it is based ; and thus the two main types of procedure tend to merge the one into the other.

III. THE TREATMENT OF MULTIPLE DETERMINATES

The General Table of Observed Data. With the Binet-Simon scale, and, indeed, with most of the qualitative tests which the psychologist is likely to factorize, the presence or absence of the 'attribute' yields in each case a dichotomous or 'bipolar' classification of the individuals observed. For the majority of such cases I have suggested that the φ-coefficient can be regarded as the simplest and most convenient procedure.[1] However, in constructing contingency-tables, it is often essential to consider cases in which the classification is (to preserve Yule's terminology) not 'twofold,' but 'manifold.'[2] Tables of this type, and particularly the compound table, where there are not only more than two possibilities for each trait, but also more than two traits (a type of table which Yule barely mentions), have been much neglected in the past ; and it is one of the chief objects of this paper to direct attention to the problems which it raises. Evidently a twofold classification may always be regarded as a special and simple case of a manifold classification. Hence it will be more instructive to discuss the question in its most general form.

In a 'manifold' classification "the individuals are first divided under n_i heads, A_1, $A_2, \ldots A_{n_i}$" ; this classification may then be compared with a number of other classifications, dividing the same individuals into the classes $B_1, B_2, \ldots B_{n_j}$, and so on (6, p. 60). To meet the more complex cases I propose to introduce a terminology familiar to logicians, and call the general character to which the classification relates (*i.e.*, what scholastic writers termed the *fundamentum divisionis*) a 'determinable,' and the alternative qualities assigned to the determinable the 'determinates.'[3] Thus, to take Yule's example, hair-colour may be called a 'determinable,' and the various shades that hair-colour is recorded as taking—'fair,' 'red,' 'dark'—the 'determinates.' Accordingly, the matrix of observed data which we require to investigate will have the form[1] shown in Table I. Let us call this matrix M_1, since it corresponds to the non-standardized matrix of observed measurements in quantitative investigations.

---

[1] The problem arose afresh in connexion with the analysis of data accumulated by psychologists working in the Directorate of Personnel Selection at the War Office during the war. Much of the present section is based on a Memorandum which I was asked to submit in 1943. For the special case in which the data yield simple dichotomous classifications I there suggested the procedure described in *Mental and Scholastic Tests*, namely, substituting a point-distribution calculated from the resulting 2 × 2 tables in place of the absolute or relative frequencies. A slightly different mode of approach was suggested by Mr. P. Slater, which he has since described more fully in a suggestive paper on 'The Factor Analysis of a Matrix of 2 × 2 Tables' (*J. Roy. Stat. Soc.*, IX (Sup.), 1947, pp. 114–27). Apart from theoretical objections, I am inclined to think that the working procedure would in practice be found somewhat laborious. Moreover, most of the traits with which he is concerned ("shyness," "apprehensiveness," "emotional instability" and the like) seem to me to imply *graded* characteristics, which (so far as is known) have an approximately normal distribution. Hence for these the tetrachoric coefficient would seem to be more appropriate.

[2] *Loc. cit.*, p. 60.

[3] I have borrowed the terms from W. E. Johnson, *Logic* (1921), Pt. I, Chap. XI, p. 174.

TABLE I. OBSERVED TABLE SHOWING PRESENCE OR ABSENCE OF ATTRIBUTES FOR A GIVEN SAMPLE OF PERSONS

	Persons (N)					Total for Row
	Tom	Dick	Harry	...	George	
Determinables (m)						
Hair-colour						
Fair	1	0	0	...	0	N_1
Red	0	1	0	...	0	N_2
Dark	0	0	1	...	1	N_3
Subtotal	1	1	1	...	1	N
Eye-colour						
Light	1	0	1	...	0	N_4
Brown	0	1	0	...	1	N_5
Subtotal	1	1	1	...	1	N
Total for Column ..	m	m	m	...	m	Nm

It will be convenient to adopt the following notation :[2]

Number of determinables	m,
Number of determinate values in the 1st, 2nd, . . . *j*th, . . . *m*th determinable	$n_1, n_2 \ldots n_j, \ldots, n_m$,
Total number of determinates	$n = n_1 + n_2 + \ldots + n_m$
Total number of persons for each determinate ..	$N_1, N_2, \ldots N_{nj}$,
Total number of persons in sample	$N = N_1 + N_2 + \ldots + N_{nj}$,
Number of persons possessing both the *j*th and the *k*th determinates	N_{jk}.

Since for each determinable every person must possess one determinate value and none can possess more than one (e.g., every person must have hair of *some* colour—either fair, dark, or red), the subtotal obtained by each person for each determinable must be 1. It follows that, if for the first determinable (say) the number of rows is n_1, then, when we know the values for $(n_1 - 1)$ of those rows and knowing the values for the row of totals (I), we can immediately deduce the value for the remaining row. Accordingly, assuming $N > n$, the rank of the matrix represented by Table I will be $(n_1 - 1) + (n_2 - 1) + \ldots + (n_m - 1) + 1 = n - m + 1$.

---

[1] Since for each determinable the last row of frequencies can be deduced from the preceding rows, we may, if we prefer, omit it from the subsequent calculations : that would evidently give rise to a slightly different form of analysis, but the modifications entailed will be obvious. When every one of the determinables permits of two alternatives only (as in dealing with characteristics recorded as 'present' or 'absent,' or with test-items that are marked only 'pass' or 'fail'), and when we record only one of the two determinates, the table reduces to the form originally used for the Binet tests. Let me add that, in certain cases, it is desirable to work with compound characteristics. Thus not infrequently the mere presence of one or other of two characteristics may have no diagnostic significance by themselves, but their joint presence may be highly significant. For example, in an adolescent girl, pallor by itself has little importance, and increasing plumpness still less ; but the two together will at once be suggestive of chlorosis. Similarly in emotional and moral disturbances, it is the total syndrome that is of diagnostic importance, not the isolated symptom. For statistical purposes, any such compound characteristic should still be treated as a single attribute, even though it involves a cluster of traits.

[2] Here and later, instead of Yule's notation for contingency-tables, I shall keep, so far as possible, to that introduced by Kelley and adopted by most statistical psychologists in this country (cf. Kelley, T. L., *Statistical Method*, 1923, pp. 263f.).

Since $f'(R - v\,I) = 0$, we must have $f'R = v f'$: i.e., if f' denotes a row of saturations for a principal axis of the symmetrical matrix R, the row of weighted column sums obtained by post-multiplying the saturations by the matrix R must be proportional to the original row of saturations, and the constant of proportionality must be $f'f = v$, the factor-variance or 'latent root.' Accordingly, consider the section of the *i*th column of R for the first determinable. This has the following form :

$$\begin{matrix} \dfrac{N_{1j}}{\sqrt{N_1\,N_j}} \\ \dfrac{N_{2j}}{\sqrt{N_2\,N_j}} \\ \dfrac{N_{3j}}{\sqrt{N_3\,N_j}} \\ \cdots \end{matrix}$$

On multiplying each row of R by the saturation coefficient, we obtain

$$\begin{matrix} \dfrac{N_{1j}}{\sqrt{N_1\,N_j}} \cdot \sqrt{\dfrac{N_1}{N}} & = & \dfrac{N_{1j}}{\sqrt{N_j\,N}} \\ \dfrac{N_{2j}}{\sqrt{N_2\,N_j}} \cdot \sqrt{\dfrac{N_2}{N}} & = & \dfrac{N_{2j}}{\sqrt{N_j\,N}} \\ \dfrac{N_{3j}}{\sqrt{N_3\,N_j}} \cdot \sqrt{\dfrac{N_3}{N}} & = & \dfrac{N_{3j}}{\sqrt{N_j\,N}} \\ \cdots & & \cdots \end{matrix}$$

But the total number showing these determinate values is N_j (i.e., if there are only three such values, $N_{1j} + N_{2j} + N_{3j} = N_j$). Hence the total of the column will be $\dfrac{N_j}{\sqrt{N_j\,N}} = \sqrt{\dfrac{N_j}{N}}$. If there are m determinates, the total for each column, after every row has been weighted by the appropriate saturation coefficient, will be $m\sqrt{\dfrac{N_j}{N}}$. Dividing by m (the latent root), we obtain the saturation coefficient. The required condition is thus fulfilled.

Since the number of factors in R (i.e., the rank of R) is less than the number of determinate traits (i.e., the order of R), each of the remaining latent roots will have a value that is less than m. Hence the factor just extracted is the factor which accounts for the maximum amount of the total variance, and may therefore be regarded as the first or general factor.

From the values given for the saturations by equation xii, we can now reconstruct the symmetrical hierarchy due to the first factor, namely, $H = ff'$. We obtain the symmetrical matrix

$$\begin{bmatrix} \dfrac{N_1}{N} & \dfrac{\sqrt{N_1 N_2}}{N} & \cdots & \dfrac{\sqrt{N_1 N_n}}{N} \\ \dfrac{\sqrt{N_2 N_1}}{N} & \dfrac{N_2}{N} & \cdots & \dfrac{\sqrt{N_2 N_n}}{N} \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

We may write the general element in this matrix $h_{ij} = \dfrac{\sqrt{N_i N_j}}{N}$. But the general element in the standardized product-sum matrix was $r_{ij} = \dfrac{N_{ij}}{\sqrt{N_i N_j}}$. Hence the general element in the residual matrix will be

$$\begin{aligned} r_{ij} - h_{ij} &= \frac{N_{ij}}{\sqrt{N_i N_j}} - \frac{\sqrt{N_i N_j}}{N} \\ &= \frac{1}{\sqrt{N_i N_j}} \left(N_{ij} - \frac{N_i\,N_j}{N}\right). \end{aligned}$$

Now $\dfrac{N_i N_j}{N}$ is the 'expected value' of the frequency in the *ij*th subgroup. This 'expected value' is the frequency we should deduce for that subgroup if it consisted simply of individuals sampled at random from the general population represented by the N persons whom we have observed, and if there were no special correlation between the ith and the jth traits, due to some causal influence affecting those traits. Using a notation that has recently come into popular use [1] let us write $\frac{N_i N_j}{N} = E$ for the expected frequency, and $N_{ij} = O$ for the frequency actually observed. Then $N\,(r_{ij} - h_{ij})^2 = \frac{(E - O)^2}{E}$; and thus expresses the square of the discrepancy between the observed value and the expected value for any element in the product-sum matrix. Now, if such discrepancies were due solely to chance, we should expect the sum of their squares to be distributed in accordance with chi-squared. In fact, $\chi^2 = \Sigma \frac{(E - O)^2}{E}$ is the value we should calculate if we wished to test the set of observed data to determine whether or not they represented the results of random sampling from the general population.[2]

The meaning of what we have called the saturations for the first factor is obvious. In factor analysis the first or 'general' factor describes the general quality defining the entire *genus* with which we are dealing. Its saturations express the relative importance of the several traits as manifestations of that generic factor. Here they are actually proportional to $\sqrt{N_1}, \sqrt{N_2}, \ldots \sqrt{N_n}$, i.e., to the square root of the frequency with which the several traits occur in the population sampled. The factor-measurements for each individual can be calculated (if required) in accordance with the usual procedure. They will evidently show how far that person is typical of the general population represented by the sample.

The factor-saturations for the remaining factors can be extracted in the usual way and tested by the chi-squared procedure.

The Conditions of Consistency. Figures descriptive of one and the same population or sample must be mutually compatible, and therefore obey certain restrictive conditions. The precise formulation of such conditions has a double value: it enables us (i) to detect data that are inaccurate and (ii) to assign limits to the unknown values when the data are incomplete. Elsewhere I have ventured to suggest that, whether the traits concerned are qualitative attributes or quantitative variables,[3] the most general criterion is essentially the same, namely, that *each of the principal minors of the symmetric table of product-sums shall be non-negative.*

In *Laboratory Notes on Factor Analysis* this principle was deduced for quantitative variables; it remains to show that it also holds good for qualitative data. The arguments are much the same in either case. Let R_i denote a symmetric submatrix, obtained from R (eq. iii) by striking out corresponding rows and columns for certain variables. Then $|R_i|$ will be a principal minor of R. Now R_i, like R, can be factorized into its latent vectors and roots. Thus, we can write $R_i = L_i V_i L_i'$, and therefore $|R_i| = |L_i||V_i||L_i'| = |V_i|$ (since L_i is orthogonal and $L_i = 1$). When the rank of R_i (r_i, say) is equal to its order (n_i), then $V_i = F_i' F_i$ and forms a diagonal matrix in which each element consists of the sum of the *squares* of the saturations. When $r_i < n_i$, the extra elements in V_i will be zero. Hence, whatever the rank of R, no element in V_i can be negative; therefore the value of $|R_i|$ is also non-negative. This clearly holds for *any* principal minor in R, and consequently for any that can be derived from C.

In practice it is unnecessary to test *every* principal minor. By invoking the Binet-Cauchy theorem, it can readily be proved that, if the *leading* principal minors of the 1st, 2nd, . . . rth order are positive, then *all* the principal minors of those orders will be positive. Now the former are the minors which automatically appear, stage by stage, when we adopt what I have called the 'factorial method' of calculating partial correlations by hierarchical subtraction.[4] And it is easy to show that,

---

[1] E.g., by McNemar (*Psychological Statistics*, 1949, pp. 179, 186), and other recent text-books.

[2] It may be noted that here as in an analysis of variance the averages (means) are not eliminated from the very outset as though they were devoid of interest, and as they would be in an ordinary correlational procedure. They are, in effect, tested for significance.

[3] In the case of *qualitative attributes*, the working criterion suggested by Yule is that no ultimate class frequency can be negative; in the case of *quantitative variables* his basic principle is that no partial correlation can exceed the numerical value of unity: cf. (6), Chap. II, pp. 17–24, 250–2, and refs.; also *id.*, 'On the Theory of Consistence of Logical Class Frequencies,' *Phil. Trans. A*, CXCVII, 1901, pp. 91f.

[4] See *Laboratory Notes*, sect. 8, 'The Solution of Simultaneous Equations': (a brief summary of the method, with a worked example, will be found in Vernon's *Notes on Statistical Method*, pp. 42f.).

The Unstandardized Product-sum Matrix. By post-multiplying the table of observed frequencies by its transpose, we obtain a product-sum matrix

$$C_t = M_1 M_1' . \qquad \text{(i)}$$

Except for the fact that the values in M_1 are not standardized deviates, this product-sum matrix corresponds with the more familiar matrix of covariances (or rather codeviances) between *traits*. We can obtain an alternative product-sum matrix by pre-multiplying M_1 by its transpose, namely, $C_p = M_1' M_1$. The result will correspond to the matrix of covariances or codeviances between *persons*. This second type of matrix, however, will not concern us at this stage.

The matrix of product-sums for traits constitutes a symmetrical contingency-table. Consider, for example, the calculation of the product-sum for a pair of determinates from two different determinables, such as those set out below :

																Total
Hair, fair (j)	1	1	1	1	0	1	0	1	1	0	1	0	1	1	0	N_j
Eyes, light (k)	1	1	1	0	1	1	1	0	1	0	0	1	1	1	1	N_k
Product..	1	1	1	0	0	1	0	0	1	0	0	0	1	1	0	N_{jk}

Evidently the product-sum, $N_{jk} = 7$, states the number of persons who possess *both* traits (a 'second order' frequency) ; and this is the value that would be entered in one cell of a small contingency-table that might be drawn up to show the relation between hair-colour and eye-colour. (For an illustration of the type of table eventually obtained, see p. 177.)

However, as the size of the sample (N) will necessarily differ from one investigation to another, it will be better to express the data in terms not of the absolute but of the relative frequencies, e.g., $p_{jk} = N_{jk}/N$ say. We can then go on to standardize the data by dividing this figure by the standard deviation, $\sqrt{N_j N_k / N}$. In this way we get a set of proportionate values which we may designate

$$r_{jk} = p_{jk} / \sqrt{p_j p_k} = N_{jk} / \sqrt{N_j N_k} . \qquad \text{(ii)}$$

This is equivalent to pre- and post-multiplying the product-sum matrix by a diagonal matrix $D^{-\frac{1}{2}}$, whose elements are $1 / \sqrt{N_j}$. We thus arrive at a modified symmetrical matrix

$$R_t = D^{-\frac{1}{2}} C_t D^{-\frac{1}{2}} = D^{-\frac{1}{2}} M_1 M_1' D^{-\frac{1}{2}} = MM' \text{ say} . \qquad \text{(iii)}$$

As will be shown in a moment, this matrix will be positive semidefinite ; being symmetrical, it will therefore have Gramian properties.[1] And further we may note that each element in its leading diagonal now has the value of unity. The matrix may thus be regarded as in standard form. (See Table III, p. 178.)

Factor Analysis. We may now proceed to factorize this standardized symmetrical matrix by finding its latent roots and latent vectors. This will enable us to express our $n \times n$ observed matrix in terms of a set of r orthogonal factors (where $r = n - m + 1$). Adopting the usual notation we shall have

$$D^{-\frac{1}{2}} M_1 = M = LV^{\frac{1}{2}} P = FP , \qquad \text{(iv)}$$

where P will be an $r \times N$ section of an orthogonal matrix ($PP' = I$), L an $n \times r$ section of another orthogonal matrix ($L'L = I$), and $V^{\frac{1}{2}}$ an $r \times r$ diagonal matrix. This gives

$$R_t = MM' = LV^{\frac{1}{2}}P \, . \, P'V^{\frac{1}{2}}L' = LVL' , \qquad \text{(v)}$$

or (on taking product-moments for persons rather than traits)

$$R_p = M'M = P'V^{\frac{1}{2}}L' \, . \, LV^{\frac{1}{2}}P = P'VP . \qquad \text{(vi)}$$

---

[1] If all the principal minors of a matrix are non-negative, i.e., greater than or equal to zero, it is said to be positive definite (or semidefinite, if its rank is less than its order). If in addition it is symmetrical, it is said to be Gramian. A related but less general property, which characterizes many contingency-tables of the kind I have described, is that which Yule terms 'isotropy' ; its presence has an obvious interest for the factorist : (see 6, pp. 68–70, and refs., p. 73).

In either case V will be the diagonal matrix containing latent roots of both R_t and R_p, and L and P will be the latent vectors of R_t and R_p respectively. To obtain appropriate equations for calculating P from M, say, $W'M = P$, we have $L'M = L'LV^{\frac{1}{2}}P = V^{\frac{1}{2}}P$.

Hence
$$P = V^{-\frac{1}{2}} L'M\,; \tag{vii}$$
This suggests that we may put $W' = V^{-\frac{1}{2}} L' = V^{-1} F$.

Values for V and F can be computed by following the usual procedure. By hypothesis the first row of P will be a set of hypothetical measurements, one for each person, and will be so calculated that, when multiplied by the first column of the weighting matrix F, the product will be an $n \times m$ hierarchical matrix giving the closest possible fit to M. Let f be the first column of the weighting matrix. The familiar method for solving the equations to obtain f and ν leads to the introduction of ν as a Lagrange multiplier. The conditions for minimizing the squared discrepancies then yield an equation of the well-known form
$$f'(R_t - \nu I) = 0\,, \tag{viii}$$
which will only be satisfied if
$$|\, R_t - \nu I \,| = 0\,. \tag{ix}$$
The solution yields r values for ν. To obtain the first factor, we take the largest value for ν, insert it in equation (viii), and solve for f'. Then, taking the other values for ν in order of magnitude, we find the remaining rows for F'. In practice the values can be calculated by the iterative method known as 'weighted summation.'

Interpretation of Factorial Results. It will be seen that the calculations I have described closely resemble those proposed by Pearson for analysing a correlated set of *quantitative* traits, such as height, arm-length, leg-length, etc. But, since our starting point has been a table of frequencies rather than a table of measurements, we cannot press the analogy too far without first considering in greater detail what precisely is the nature of the results thus reached.

Equation (iv) may be interpreted by saying that the modified matrix of observed data ($M = D^{-\frac{1}{2}} M_1$) has been expressed in terms of two factor-matrices: (*a*) a matrix of 'uncorrelated' factor-measurements for *persons*, P, so chosen that the first factor will account for the maximum amount of the variance exhibited by the initial matrix, and each successive factor will account for a maximum amount of the variance exhibited by each successive residual matrix; and (*b*) a matrix of 'uncorrelated' factor-saturations for the several *traits*, giving the weights to be applied to each factor in reconstructing the observed trait-measurements. If we prefer to consider ourselves as factorizing the unmodified matrix of observed data, F, we can write equation (iv) in the form $M_1 = (D^{\frac{1}{2}}F)P$, and $D^{\frac{1}{2}}F = D^{\frac{1}{2}}LV^{\frac{1}{2}}$ will now represent the factor-saturations or weights; but the columns of factor-saturations ($D^{\frac{1}{2}}F$) will no longer be uncorrelated.

Moreover, if in equation (viii) we substitute for R_t from equation (iii), we obtain
$$f'(D^{-\frac{1}{2}} M_1 M_1' D^{-\frac{1}{2}} - \nu I) = 0\,. \tag{x}$$
We can simplify this by writing $f' D^{-\frac{1}{2}} = u'$ and post-multiplying by $D^{\frac{1}{2}}$. We thus obtain
$$u'(M_1 M_1' - \nu D) = 0, \text{ that is, } u' M_1 M_1' = \nu u'D\,. \tag{xi}$$
We now see that one solution of this equation will be obtained by taking $u' = [1, 1, \ldots 1] = w_0'$ (in the usual notation), and $\nu = m$. For $w_0' M_1 = [m, m, \ldots m]$ (i.e., the sums of the several columns in Table I) $= m\, w_0'$; and $w_0' M_1' = [N_1, N_2 \ldots N_n]$ (i.e., the sums of the several rows of the same table) $= w_0' D$. We thus obtain
$$u' M_1 M_1' = m\, w_0' M_1' = m\, w_0' D = \nu u' D\,,$$
in accordance with equation (xi). This procedure will yield only the proportionate values for the latent vectors, in this case $f' = w_0' D^{\frac{1}{2}} = [\sqrt{N_1}, \sqrt{N_2}, \ldots \sqrt{N_n}]$. But since the elements of l' are also proportional to those of f', and $l' l = 1$, we can first calculate values of l' by normalizing the above values for f'. This gives
$$l' = \left[\sqrt{\frac{N_1}{N_m}}, \sqrt{\frac{N_2}{N_m}}, \ldots \sqrt{\frac{N_n}{N_m}}\right], \text{ since } N_1 + N_2 + \ldots + N_n = N_m\,.$$
Now $f' = \nu^{\frac{1}{2}} l' = \sqrt{m}\,.\,l'$. Hence we can at length obtain appropriate values for the saturations namely,
$$f' = \left[\sqrt{\frac{N_1}{N}}, \sqrt{\frac{N_2}{N}}, \ldots \sqrt{\frac{N_n}{N}}\right]\,. \tag{xii}$$

It is easy to show algebraically that the expression thus reached for the saturations conforms with the requirements for a first 'principal axis.' The requirements can be formulated quite simply.

if the condition specified is fulfilled, then (i) in the case of qualitative data all frequencies up to those of the order included in the product-sum matrix must have real and positive values, (ii) in the case of quantitative data no partial correlation can exceed the numerical value of 1, and (iii) in either case the tables can be analysed in terms of real factors. It will be sufficient to indicate quite briefly the general lines of proof for these three statements.

(i) *Higher Frequencies.* Since the square-sums of frequencies are necessarily larger than their product-sums, the appearance of a negative principal minor would imply that some of the frequencies had negative or imaginary values. If the product-sum matrix includes only frequencies of the second order, the condition as stated is necessary but not sufficient. To show that the *ultimate* frequencies have real and positive values, it would be necessary to introduce composite attributes.

(ii) *Partial Correlations.* No generality will be lost if we suppose the correlation matrix or product-sum to be arranged so that the rows and columns for the two correlated variables shall be the last. Any variables which are neither to be correlated nor to be partialed out will be deleted from the matrix. Then, the partial correlation,

$$r_{(n-1)n.12\ldots(n-2)} = r \text{ (say)} = \frac{D_{(n-1)n}}{\sqrt{D_{(n-1)(n-1)}}\ \sqrt{D_{nn}}}, \quad . \quad . \quad . \quad \text{(xiii)}$$

where D, the determinant obtained from R after the irrelevant rows and columns have been deleted, will be a principal minor of R. Now, if $1 \geqslant r^2$, evidently $D_{(n-1)(n-1)}\, D_{nn}$ must be $\geqslant D^2{}_{(n-1)n}$; that is

$$\begin{vmatrix} D_{(n-1)(n-1)}, & D_{(n-1)n} \\ D_{(n-1)n} & , D_{nn} \end{vmatrix} = \text{non-negative.} \quad . \quad . \quad . \quad \text{(xiv)}$$

Now

$$D = \begin{vmatrix} D_{n(n-1),n(n-1)} \end{vmatrix}^{-1} \begin{vmatrix} D_{(n-1)(n-1)}, & D_{(n-1)n} \\ D_{(n-1)n} & , D_{nn} \end{vmatrix}, \quad . \quad . \quad . \quad \text{(xv)}$$

where the divisor must be positive. Hence, if D (the principal minor) is non-negative, then the expression on the left of (xiv) must be non-negative, and we have $r^2 \leqslant 1$; if D is negative, then $r^2 > 1$.

(iii) *Real Factors.* Now, whatever factorial methods we adopt to reduce R to a diagonal matrix, V say, we can split this diagonal matrix, and so further reduce R to the form $R = F\,I\,F'$, where $F = LV^{\frac{1}{2}}$. But this further process will in general be non-rational[1]; and if any of the non-zero elements in V (and therefore in I) are negative, $V^{\frac{1}{2}}$ and consequently F will involve non-real elements. But, by Sylvester's law of inertia, however we reduce R to the form $F\,I\,F'$, the number of negative values in I must remain the same. Therefore, if either of the methods of factorization described above yields imaginary elements for the factor matrix F, then *any* method must yield such elements: that is, it will be impossible to analyse R into factors with real elements. Conversely, if our criterion is fulfilled, then R will be reducible to terms of real factors.

We may therefore sum up our conclusions regarding the table of product-sums as follows: in order that an $n \times n$ table shall be (*a*) self-consistent and (*b*) analysable in terms of r *real* factors ($r \leqslant n$), it is at once necessary and sufficient that *all its leading principal minors shall be non-negative.*

IV. A FACTORIAL ANALYSIS OF PHYSICAL ATTRIBUTES

The Neglect of Qualitative Traits in Studies of Body Types. To illustrate the general procedure proposed may I now describe its application in a simple concrete case? The various points in my argument will perhaps be most readily grasped if I choose a problem from physical anthropology. The characters in question will then be easier to visualize; less doubt is likely to arise about which characters are genuinely qualitative; and at the same time it will be possible to show how quantitative measurements may be included in the same series and subjected to the same general treatment.

Recent discussions about the relation between temperamental and physical characteristics have been limited almost entirely to those physical characteristics which, like height and weight, may be treated as continuously graded, quantitative variables; they have tended almost entirely to overlook the possible relevance of qualitative characteristics, such as eye-colour, hair-colour, or peculiarities in the shape of certain features, like the nose or chin (cf. 10 and refs.). In a later paper I hope to examine what verifiable support there may be for popular and traditional notions about the value of such traits as a clue to individual character. Here my purpose is primarily to illustrate a practicable technique for expressing, in terms of 'factors' or type-tendencies, any significant associations that may be observable among the physical traits themselves.

---

[1] Cullis, C. E., *Matrices and Determinoids*, II, p. 293, Theorem III, and *Factors of the Mind*, p. 307, footnote 1.

The Data. In an early investigation which I commenced some years ago, I endeavoured to determine more particularly whether there was any evidence for such associations in the case of those traits that were commonly supposed to characterize certain 'racial' types (cf. 7, pp. 23, 27, and refs.). For this purpose I collected observations on the characteristics most frequently used in classifying individuals into such groups, namely, eye-colour, hair-colour, stature, shape of head, shape of nose and chin, and type of hair. Here it will be unnecessary to defend or define the various terms adopted: as regards stature it will be sufficient to say that persons were classed as 'tall' if their height was above the average for the general population [1] and 'short' if it was below the average. Under 'wide' heads I here include all with a cranial index over 75 (that is, mesocephalic heads, as well as dolichocephalic in the stricter sense). To shorten the table, I have also pooled certain traits and omitted others.

The data were collected at Liverpool, a district where representatives of different nationalities—Welsh, Irish, Scots, as well as foreigners—were easily accessible. In most cases temperamental assessments were obtained at the same time; but these will also be omitted from the present tables, as they raise somewhat different issues. In all, 217 individuals were examined, about two-thirds of them males. But, partly to simplify the calculations and partly because the later observations were rather more trustworthy, I shall here restrict my analysis to the data obtained from the last hundred males in the series.

The Crude and Standardized Contingency-Tables. The number of persons characterized by the several attributes specified, taken in pairs, are set out in Table II (this may be taken as representing the initial type of contingency-table which was designated C_t on p. 172 above).

TABLE II. OBSERVED FREQUENCIES

Number of Persons exhibiting Characteristics Specified

Trait	F	R	D	Tot.	L	M	B	Tot.	N	W	Tot.	T	S	Tot.
HAIR														
Fair ..	22	0	0	22	14	6	2	22	14	8	22	13	9	22
Red ..	0	15	0	15	8	5	2	15	11	4	15	10	5	15
Dark ..	0	0	63	63	11	25	27	63	44	19	63	20	43	63
Total ..	22	15	63	100	33	36	31	100	69	31	100	43	57	100
EYES														
Light ..	14	8	11	33	33	0	0	33	27	6	33	29	4	33
Mixed ..	6	5	25	36	0	36	0	36	20	16	36	10	26	36
Brown ..	2	2	27	31	0	0	31	31	22	9	31	4	27	31
Total ..	22	15	63	100	33	36	31	100	69	31	100	43	57	100
HEAD														
Narrow..	14	11	44	69	27	20	22	69	69	0	69	30	39	69
Wide ..	8	4	19	31	6	16	9	31	0	31	31	13	18	31
Total ..	22	15	63	100	33	36	31	100	69	31	100	43	57	100
STATURE														
Tall ..	13	10	20	43	29	10	4	43	30	13	43	43	0	43
Short ..	9	5	43	57	4	26	27	57	39	18	57	0	57	57
Total ..	22	15	63	100	33	36	31	100	69	31	100	43	57	100

---

[1] By the general population I mean the population of England and Wales (the means were taken from the Anthropometric Committee of the British Association, *Brit. Ass. Ann. Report*, 1883, pp. 256f.). The definitions of the set of traits originally used also, for the most part, followed those of the same Committee. The original classification of hair- and eye-colour was somewhat more elaborate than that shown in the present tables (cf. 11, p. 150, 12 and refs.); but the sample is too small to permit of numerous fine subdivisions.

The reason for selecting the four characteristics set out in Table II will be obvious: as Dr. Morant has observed, "the classical anthropological classifications of the peoples of Europe were based almost entirely on the cephalic index, stature, and hair- and eye-colour; . . . it is not to be doubted that they make clear distinctions between the major groups in the continent as a whole." (*Races of Central Europe*, 1939, p. 135.)

We begin by dividing first the rows of Table II, and then the resulting columns, by the square root of the number in the corresponding diagonal cell (the number of persons characterized by each of the eight attributes). This converts the number in each diagonal cell to unity. The resulting figures are shown in Table III, which corresponds to the matrix R_t, p. 172. This is the matrix that we shall now proceed to factorize.

TABLE III. RELATIVE FREQUENCIES

Trait	F	R	D	L	M	B	N	W	T	S
HAIR										
Fair.. ..	1·000	·000	·000	·519	·213	·077	·359	·306	·423	·254
Red	·000	1·000	·000	·359	·215	·093	·342	·185	·394	·171
Dark ..	·000	·000	1·000	·241	·525	·611	·667	·430	·384	·718
EYES										
Light ..	·519	·359	·241	1·000	·000	·000	·566	·188	·770	·092
Mixed ..	·213	·215	·525	·000	1·000	·000	·401	·479	·254	·574
Brown ..	·077	·093	·611	·000	·000	1·000	·476	·290	·110	·642
HEAD										
Narrow ..	·359	·342	·667	·566	·401	·476	1·000	·000	·551	·622
Wide ..	·306	·185	·430	·188	·479	·290	·000	1·000	·356	·428
STATURE										
Tall	·423	·394	·384	·770	·254	·110	·551	·356	1·000	·000
Short ..	·254	·171	·718	·092	·574	·642	·622	·428	·000	1·000

Factor Analysis of Relative Frequencies. The saturations for the first factor, it will be remembered, can be determined by simply dividing the square root of the number of persons characterized by each attribute (N_i) by the total number ($N = 100$). The values so calculated are set out in the first column of Table VIIA. As may easily be verified, the factor-variance, i.e., the sum of the squares of these values, is $m = 4$. The reader may find it instructive to check the fact that the saturations correspond to the first principal component, by multiplying each row of Table III (or, say, the first two or three figures from each row) by the corresponding saturation, and dividing each weighted sum by the factor-variance, namely, 4. The results should be equal to the factor-saturations employed.

A hierarchy is now constructed by multiplying each saturation in turn by all the saturations. This is subtracted from Table III; and the residuals so obtained are given in Table IV.

TABLE IV. FIRST RESIDUALS

Trait	F	R	D	L	M	B	N	W	T	S
HAIR										
Fair ..	·780	−·182	−·372	·250	−·068	−·185	−·030	·045	·115	−·100
Red	−·182	·850	−·370	·137	−·017	−·123	·020	−·030	·140	−·121
Dark ..	−·372	−·370	·370	−·215	·049	·169	·008	−·012	−·136	·118
EYES										
Light ..	·250	·137	−·215	·670	−·345	−·320	·089	−·132	·393	−·341
Mixed ..	−·068	−·017	·049	−·345	·640	−·334	−·097	·145	−·139	·121
Brown ..	−·185	−·123	·169	−·320	−·334	·690	·013	−·020	−·255	·222
HEAD										
Narrow ..	−·030	·020	·008	·089	−·097	·013	·310	−·462	·006	−·005
Wide ..	·045	−·030	−·012	−·132	·145	−·020	−·462	·690	−·009	·008
STATURE										
Tall	·115	·140	−·136	·393	−·139	−·255	·006	−·009	·570	−·495
Short ..	−·100	−·121	·118	−·341	·121	·222	−·005	·008	−·495	·430

A minor peculiarity in the three foregoing tables deserves a word of comment. In the diagonal submatrices of Tables II and III, all the values except those in the leading diagonal are zero, and thus "all the frequency is concentrated in the diagonal compartments." [1] This is plainly due to the fact that all the alternative determinates under one and the same determinable are, by hypothesis, mutually exclusive: that being so, of the 22 persons who are entered as having fair hair none can also be entered as having red or dark hair. In the corresponding cells of Table IV the negative residuals indicate that, after allowing for the effect of the first or general factor, we find that any tendency to grow fair hair inevitably has a negative effect on the tendency to grow hair of any other colour.

The next step is to test the residuals for significance. For this purpose we must obviously confine our examination to those residuals resulting from figures that have been actually observed, i.e., those appearing in the non-diagonal matrices, relating to *different* determinables. If the reader calculates the sum of the squares of the figures in each of the side submatrices in Table IV, and then multiplies that sum by $N = 100$, he will find that the results agree exactly with the figures obtained for chi-squared when calculated in the usual way from the corresponding submatrices in Table I. The six values thus obtained are shown in Table V. All except those for the association of head-shape with hair-colour and stature are significant. The total value for chi-squared comes to 78·35, which, with 13 degrees of freedom, is highly significant.

TABLE V. CHI-SQUARE

Determinable	HAIR	EYES	HEAD	STATURE
HAIR ..	—	21·26	0·45	9·01
EYES ..	21·26	—	5·63	41·98
HEAD ..	0·45	5·63	—	·02
STATURE	9·01	41·98	·02	—

TABLE VI. POINT-CORRELATIONS

Determinable	HAIR	EYES	HEAD	STATURE
HAIR ..	1·000	·386	−·043	·284
EYES ..	·386	1·000	·195	·553
HEAD ..	−·043	·195	1·000	·014
STATURE	·284	·553	·014	1·000

The residuals in Table IV may be regarded as expressing partial or specific associations. They can be factorized in their turn; and the process repeated. Since there are only seven degrees of freedom, only seven factors can be extracted. All except the last are fully significant. The saturations [2] are given in Table VIIA. Most of the factors can readily be interpreted. The first is a general factor, i.e., all its signs are positive: as we have already seen, it specifies the general pattern of traits (or trait-tendencies) which characterizes the particular sample of the British population we have been examining: in the main they tend to be dark-haired, of mixed eye-colour, dolichocephalic, and short rather more frequently than tall. The second factor is bipolar: it indicates a contrast between (*a*) a complex tendency which combines light eyes with tall stature, fair or red hair, and a narrow head, and (*b*) an opposite tendency to combine short stature with dark hair, brown or (less frequently) mixed eye-colour, and a broad head: of the two antithetical types light eyes in the one case and short stature in the other appear to be the traits which are most diagnostic. The third and fifth factors appear to contrast a brachycephalic tendency (and its associated characteristics) with a dolichocephalic; the fourth to contrast the red-haired type with the fair-haired; and the sixth suggests a minor type combining tall stature with dark hair and light or mixed eye-colour.

I do not suggest that the particular analysis given is either a final analysis or the best analysis. For one thing, several of the factor-saturations, e.g., those for fair and red hair in factors III and IV, suggest that (if only a slightly larger number was available) a group factor analysis or a method of subdivided factors would yield a clearer picture. Further, the large saturations (over 0·6) that stand in isolation—for eye-colour in factor II, head-shape in factor III, and hair-colour in factor IV—tempt us to ask whether we might not secure a truer result if, instead of retaining 1·000 in the leading diagonal, we substituted a reduced figure which would exclude specific factors. But to discuss these further issues would take us too far from our immediate task.

---

[1] The description is quoted from Yule, *loc. cit.*, p. 65.

[2] The factor-saturations have been calculated by the method of weighted summation. They therefore represent 'principal components.' Accordingly, the reader, if he wishes, can apply Bartlett's test for significance (cf. this *Journal*, III, pp. 78, 118). But in cases like the present I should consider the chi-squared test more appropriate, since this can be calculated for particular submatrices only.

I should add that I am deeply indebted to Mr. Arthur Summerfield for assistance with the computations.

TABLE VII. BIPOLAR FACTOR SATURATIONS

A. Analysis of Relative Frequencies

Traits	Factors						
	I	II	III	IV	V	VI	VII
HAIR							
Fair	·469	·420	·231	−·602	−·100	−·404	·119
Red	·387	·333	−·019	·751	·182	−·376	·012
Dark	·794	−·411	−·128	−·011	−·030	·422	−·076
EYES							
Light	·574	·719	−·164	−·104	·042	·084	−·325
Mixed	·600	−·241	·563	·227	−·442	·048	·125
Brown	·557	−·482	−·437	−·138	·443	−·138	·201
HEAD							
Narrow	·831	·081	−·428	·039	−·338	−·044	·056
Wide	·557	−·121	·638	−·058	·504	·065	−·084
STATURE							
Tall	·656	·619	·010	·047	·120	·335	·241
Short	·755	−·537	−·009	−·041	−·104	−·292	−·209

B. Analysis of Point-Correlations

Traits	Factors			
	A	B	C	D
HAIR	·625	·508	·429	−·410
EYES	·820	−·176	−·478	−·410
HEAD	·448	−·709	·478	·410
STATURE	·711	·378	−·429	·410

Factor Analysis of Point-Correlations. As we have already seen, in dealing with qualitative attributes of a dichotomous type, i.e., in cases where each determinable furnishes only *two* mutually exclusive determinates, a set of product-moment correlations derived from fourfold tables will commonly provide the simplest basis for factor analysis. But it is, I think, easy to show that for data of the present type, where several of the determinables lead to a multiple classification, this method is far from satisfactory. However, let us briefly examine its application here.

In Table II, one of the submatrices (that for Head and Stature) involves a single degree of freedom only. For this pair of variables we can at once combine the four figures there given into a single figure expressing the root-mean-square contingency ($\varphi = \sqrt{\chi^2/N}$). This may be regarded as giving the product-moment correlation for the two determinables on the assumption that the frequencies for the four determinates have a point-distribution. Where the submatrices contain more than four figures, we can, if we wish, calculate a coefficient of contingency by Pearson's formula, or better, allow for the increased degrees of freedom, by dividing the corresponding φ by the fourth root of the degrees of freedom.[1] But it is better still to pool the initial data so as to secure a fourfold table for every pair of determinables from the outset. In view of the results just obtained we may plausibly group fair and red hair together; for eye-colour the original records make it possible to divide most of those classed above as 'mixed' either with the 'light' or with the 'brown eyes' respectively, and the rest may be divided in due proportion.[2] We can then calculate point-correlations for the resulting fourfold tables. In doing so, I have taken what we may call Nordic traits (fair hair, light eyes, narrow head, and tall stature) to be positive qualities. The coefficients thus obtained are shown in Table VI.

[1] This expression has been proposed by A. A. Tschuprow (8). It appears to have been suggested by the 'divergence coefficient' of Lexis (1): see also 13, p. 325. Here the coefficients so calculated are in moderately close agreement with the φ-coefficients set out in Table VI.

[2] It might perhaps be argued that we could preserve the distinction between fair hair and red hair by constructing *two* sets of fourfold tables for hair-colour, based on the dichotomies (i) 'fair' or 'not-fair (i.e., red and dark)' and (ii) 'red' or 'not-red (i.e., fair and dark).' But both logical considerations and the empirical results make it clear that this alternative procedure only increases the confusion when we come to interpret the factor-saturations so obtained.

These coefficients can readily be factorized; and the saturations, calculated by simple summation, are shown in Table VIIB. It will be seen at once that, without the more detailed knowledge reached by the method previously described, it would be very difficult to arrive at a convincing interpretation. The first factor (A) obtained from the table of point-correlations obviously corresponds, as we should expect, with the second factor (II) obtained from the table of relative frequencies: thus factor A expresses as a general factor what factor II expressed as a bipolar factor, namely, the tendency of the four positive (Nordic) traits to go together. Factor D expresses the contrast between (*a*) the combination of dark hair and dark eyes and (*b*) wide head and short stature shown by the saturations for these qualities in factor III. Factor C expresses the contrast between (*a*) the combination of dark hair and wide head and (*b*) the combination of brown eyes and short stature suggested by factor IV. Factor B apparently expresses the contrast between (*a*) the combination of red hair and tall stature and (*c*) light eyes and narrow head expressed by factor IV. However, with only four determinables, and with unity in the leading diagonal, it might well be maintained that the last three factors mainly represent specific factors expressed in bipolar form (see 10, pp. 112f.).

It must therefore be admitted that cases of manifold classification—i.e., determinables involving more than two determinates—cannot as a rule be satisfactorily treated by forcing the multiple determinates into twofold form and then analysing the coefficients derived from the resulting fourfold tables.

TABLE VIII. GROUP FACTOR SATURATIONS

Analysis of Point Correlations

Traits	Factors			
	A	B	C	D
HAIR ..	—·043	·999	·000	·000
EYES ..	·195	·395	·457	·773
HEAD ..	1·000	·000	·000	·000
STATURE	·014	·285	·958	·000

It is always instructive to supplement a bipolar analysis by a group factor analysis. The patterns exhibited by Tables VI and VIIB, suggest that the most appropriate procedure will be that which I have described elsewhere for cases in which the variables reveal a 'cumulative complexity.'[1] The result is a matrix of 'completely overlapping' (as distinct from 'partly overlapping' or 'non-overlapping') group factors (Table VIII). Here the implication of the figures is that the first classification should be based predominantly on head-shape; the second on hair-colour, eye-colour, and stature, with greatest emphasis on hair-colour; the third on the joint variation of eye-colour and stature, regardless of hair-colour; and the last on divergences in eye-colour from that indicated by the preceding classifications.

The Examination of Ultimate Frequencies. Had we been factorizing a correlation table of the ordinary kind for quantitative variables, we should, after making a bipolar analysis like that shown in Table VIIA, have proceeded at once to take the bipolar factors as indicating the way in which the several traits might be combined into groups, and then use this classification as the basis for the analysis into group factors. With qualitative traits a more direct procedure is available for determining what groupings are most significant.

Following Yule's terminology, and modifying his definition slightly so as to include those cases where the classifications may be manifold, let us term the classes specified by *all* the determinates that could be observed in any case, i.e., the classes of the *m*th order in the case of *m* determinables, the 'ultimate frequencies.'[2] As Yule points out, "the ultimate frequencies form a natural set in terms of which the data are *completely* given " (*ibid.*, p. 13). With a large number of determinates (such as may arise in certain types of test) it would be exceedingly laborious to examine all the ultimate frequencies in turn. However, in such cases the preliminary bipolar analysis would help us to pick out those sets of traits which go together most closely and are likely to form significant mth order combinations. In the present case both m and the n_i's are small numbers ; and consequently the *total* number of mth order combinations, namely, $n_1 \times n_2 \times n_3 \times n_4 = 3 \times 3 \times 2 \times 2$ amounts to 36 only. Here therefore it will be as simple as it is instructive to set out *all* the 4th order combinations, and examine the significance of each one.

---

[1] *Factors of the Mind*, p. 307. As there pointed out, the factor-matrix so obtained has the form of a triangular matrix, as will here be evident if the traits are re-arranged in the order described in the text (Head, Hair, Stature, Eyes). The implications suggested are illustrative only: to have any genuine value they should be based on a far larger number of traits.

[2] The definition, as phrased by Yule (6, p. 11), specifies 'attributes' only; he does not distinguish attributes which are determinates from those which are determinables. Such a distinction is not required by the types of problem with which he chiefly deals, namely, dichotomous classes only (presence or absence) in the case of many determinables, and two determinables only in the case of manifold determinates.

In Table IX (last column but one) I give the frequencies for each of the 36 composite categories. The number expected by chance is shown in the last column. The probability

TABLE IX. THE DETERMINATION OF SIGNIFICANT GROUPS

HAIR	EYES	HEAD	STATURE	ULTIMATE FREQUENCIES Obs.	Exp.
Fair	Light	Narrow	Tall	8	2·1
,,	,,	,,	Short	4	2·8
,,	,,	Wide	Tall	2	1·0
,,	,,	,,	Short	0	1·3
	Total			14	7·3
,,	Mixed	Narrow	Tall	1	2·3
,,	,,	,,	Short	1	3·1
,,	,,	Wide	Tall	2	1·1
,,	,,	,,	Short	2	1·4
	Total			6	7·9
,,	Brown	Narrow	Tall	0	2·0
,,	,,	,,	Short	0	2·7
,,	,,	Wide	Tall	0	0·9
,,	,,	,,	Short	2	1·2
	Total			2	6·8
TOTAL				22	22·0
Red	Light	Narrow	Tall	6	1·5
,,	,,	,,	Short	0	1·9
,,	,,	Wide	Tall	2	0·7
,,	,,	,,	Short	0	0·9
	Total			8	5·0
,,	Mixed	Narrow	Tall	2	1·6
,,	,,	,,	Short	1	2·1
,,	,,	Wide	Tall	0	0·7
,,	,,	,,	Short	2	1·0
	Total			5	5·4
,,	Brown	Narrow	Tall	0	1·4
,,	,,	,,	Short	2	1·8
,,	,,	Wide	Tall	0	0·6
,,	,,	,,	Short	0	0·8
	Total			2	4·6
TOTAL				15	15·0

TABLE IX (*continued*).

HAIR	EYES	HEAD	STATURE	ULTIMATE FREQUENCIES Obs.	Exp.
Dark	Light	Narrow	Tall	9	6·1
,,	,,	,,	Short	0	8·2
,,	,,	Wide	Tall	2	2·8
,,	,,	,,	Short	0	3·7
	Total			11	20·8
,,	Mixed	Narrow	Tall	3	6·7
,,	,,	,,	Short	12	8·9
,,	,,	Wide	Tall	2	3·1
,,	,,	,,	Short	8	4·0
	Total			25	22·7
,,	Brown	Narrow	Tall	1	5·8
,,	,,	,,	Short	19	7·7
,,	,,	Wide	Tall	3	2·6
,,	,,	,,	Short	4	3·4
	Total			27	19·5
TOTAL				63	63·0
GRAND TOTAL				100	100·0

that a given person will be at once fair-haired, light-eyed, dolichocephalic, and tall will be $\frac{22}{100} \times \frac{33}{100} \times \frac{69}{100} \times \frac{43}{100} = \cdot 0215$. Accordingly, in a sample of 100 we should anticipate that only 2·15 persons would exhibit that combination. Actually there are as many as 8. As the theoretical proportions are small we can take the standard error of the discrepancy to be approximately equal to the square root of the expected number, namely $\sqrt{2{\cdot}15}$. Hence in this case the excess over expectation is unquestionably significant.

A glance down the two columns will show that the following combinations may be regarded as fully significant :

(i) Fair hair, light eyes, narrow skull, tall stature.
(ii) Red hair, light eyes, narrow skull, tall stature.
(iii) Dark hair, brown eyes, narrow skull, short stature.

The following fall somewhat below the level of significance in the present group, but were fully significant in the larger group of 217 :

(iv) Dark hair, eyes of mixed colour, broad skull, short stature.
(v) Dark hair, light eyes, narrow skull, tall stature.

In the larger group one further combination nearly reached the level of significance, namely :

(vi) Fair hair, light or mixed eyes, broad skull, tall stature.

The first two combinations are suggestive of two familiar varieties of the so-called Nordic type ; the third is suggestive of the so-called Mediterranean type ; the fourth of the so-called Alpine or Celtic type ; the fifth is a recognized subtype usually supposed to result from a crossing of Nordic and Mediterranean types ; the sixth is suggestive of the so-called Beaker-folk. I may add that it was noted at the time that persons belonging to the first two groups frequently came from areas such as the Lake District and Yorkshire, where tall, fair-haired, light-eyed populations predominate ; however, a number of the red-haired members came from Ireland (chiefly north-eastern areas). The third group included a large proportion of Welsh or Irish immigrants ; the fourth group was predominantly Welsh, and the fifth (a very small group) predominantly Irish ; the sixth group also came chiefly from Wales. In my original analysis, place of origin was also included as a fifth determinable; but the numbers in the resulting subgroups were too small to yield significant figures. Migration and mixed marriages likewise made it clear that, with the data available, a satisfactory classification in regard to ancestral origin, which alone is really relevant, was scarcely possible.

If we regard each significant combination of traits as resting on a factor, it will be seen that in the main the results thus reached restate, in group factor form, the classification already arrived at by the bipolar analysis. It is obvious that we could pass from one type of analysis to the other by rotation; and this would yield appropriate weights for the group factors if required.

Finally, let me say that in this short paper I have only been able to touch very briefly on a few of the points raised by the problem I have been discussing and to sketch, in the barest outline, one of the many possible solutions.

V. SUMMARY AND CONCLUSIONS

1. It is argued that, if factor analysis is to be adequately employed in problems of individual psychology, a rigorous statistical treatment is needed for qualitative data no less than for quantitative; and it is claimed that, with minor modifications, factorial methods can be applied, as legitimately, as readily, and as fruitfully, with the former as with the latter.

2. The various techniques available for examining qualitative data are enumerated and briefly discussed. It is concluded that, with dichotomous classifications that are assumed to rest on a discrete point-distribution, the product-moment correlation for point-distributions (φ) will, in general, prove the most useful.

3. Cases of manifold classification, in which there are many determinables, each involving many mutually exclusive determinates, have been almost entirely ignored by previous writers. For these cases, it is shown, the point-distribution coefficient, as ordinarily calculated, is unsuitable.

4. The general methods here proposed for such cases rest on the assumption that factor analysis can be validly adopted with any type of symmetrical positive-definite matrix. The procedure suggested as appropriate for most purposes proceeds by analysing the relative frequencies. This leads to a factor matrix containing one general factor and a succession of bipolar factors. The theoretical number of factors will, as usual, depend on the rank of the matrix analysed, which will always be less than the total number of attributes. It is shown that this mode of analysis permits of employing the chi-squared criterion for testing the significance to the residuals obtained.

5. The group factor method also can be applied direct to the table of relative frequencies. But more instructive results can be obtained by examining the ultimate frequencies, or (with large tables) that particular selection of ultimate frequencies which the bipolar analysis indicates is likely to be most significant.

6. The procedure is illustrated by an analysis of the physical characteristics commonly adopted in anthropological studies of physical types, namely, hair- and eye-colour, head-shape and stature. The factors, and significant groupings found, agree in the main with those combinations alleged to be characteristic of the commoner physical types described in anthropological surveys of the British Isles. It is claimed that a factorial study of the way in which such traits are associated or combined is likely to lead to more precise and trustworthy results than the usual method of reporting averages for the separate traits.

REFERENCES

1. Lexis, W. (1886). 'Uber die Theorie der Stabilität Statischer Reihen,' *Jahrb. f. nat. Ök. u. Stat.* XXVII, 604–48
2. Yule, G. U. (1900). 'On the association of attributes in statistics,' *Phil. Trans.*, A, CXCIV, 257–84.
3. Pearson, K. (1900). 'On the correlation of characters not quantitatively measurable,' *Phil. Trans.*, A, CXCV, 1–87.
4. Pearson, K. (1904). 'On the theory of contingency and its relation to association and normal correlation,' *Drapers' Company Research Memoirs, Biometric Series*, I.
5. Lipps, G. F. (1905). 'Die Bestimmung der Abhängigkeit Zwischen den Merkmalen eines Gegenstandes,' *Ber. d. Math. phys. Klasse d. Kön. Sächs. Gesellsch. d. Wissenschaften.*
6. Yule, G. U. (1912). *Introduction to the Theory of Statistics.* 2nd ed. London : Griffin.
7. Burt, C. (1912). 'The inheritance of mental characters,' *Eugen. Rev.*, IV.
8. Tschuprow, A. A. (1918–9). 'Zur Theorie der Stabilität Statistischer Reihe,' I, 199–256, II, 80–133.
9. Burt, C. (1921). *Mental and Scholastic Tests.* London : P. S. King (2nd ed., 1946).
10. Burt, C. (1944). 'The factorial study of physical types,' *Man*, XLIV, 82–6.
11. Burt, C. (1945). 'The relation between eye-colour and defective vision,' *Eugen. Rev.*, XXXVII, 149–56.
12. Burt, C., and Grieve, J. (1945). 'Defective colour vision in relation to pigmentation and hair,' *Man*, XLV, 81–3.
13. Kelley, T. L. (1947). *Fundamentals of Statistics.* Cambridge, Mass. : Harvard University Press.
14. Burt, C. (1949). 'Alternative methods of factor analysis,' *Brit. J. Psychol., Stat. Sect.*, II, 98–121.

A NOTE ON SHELDON'S TABLE OF CORRELATIONS BETWEEN TEMPERAMENTAL TRAITS

By A. LUBIN

University College and Maudsley Hospital, London

I. *Problem.* II. *Criterion.* III. *Results.* IV. *Summary.*

I. PROBLEM

In a recent paper (1) Adcock has discussed the table of correlations between temperamental traits published by Sheldon in his book on *The Varieties of Temperament* (6), and reports considerable difficulty in finding a method of factor analysis which will account for the table as it stands. The correlations are reproduced below in Table I. As Professor Burt has pointed out, the symmetrical pattern of positive and negative signs is unlike that found in any other published table. Adcock tried various methods of factor analysis. In each case it proved impossible to complete the analysis since it led to illegitimate procedures such as taking the square root of a negative number. His own suggestion was that, to explain Sheldon's three basic variables, a complex interaction between *four* factors would have to be postulated. " The four factors " (described in the closing paragraphs of his article) " seem to lend themselves to precisely the sort of interaction which we suspected, and seem capable of explaining the unusual form of the correlations."

There is, however, a simpler hypothesis. Earlier in his paper Adcock observes : " obviously there is something peculiar about these inter-correlations." With this remark we may readily agree. And the peculiarity is so great that one is forced to ask whether it may not be outside the bounds of mathematical possibility. Hence the important question to ask is not : can these correlations be factorized ? but rather the prior question : are these correlations mutually compatible ?

In the following note, it is proposed to demonstrate that at least three of Sheldon's product-moment correlations could not be obtained simultaneously from any actual set of measurements : in short, that the table violates the well-known conditions for consistency. It follows that some at least of his figures must contain errors of arithmetical calculation.

II. CRITERION

Such errors can frequently be detected in correlation tables in virtue of certain inevitable boundary conditions that must obtain for any set of inter-correlations of three or more variables. Yule in an early paper points out that, if r_{12} and r_{13} are known, definite limits can be laid down for r_{23} : (8, p. 250 and refs. ; cf. 3, p. 387, and 5, p. 142). These may be stated in terms of r_{12} and r_{13} as follows :

$$r_{12}\, r_{13} - \sqrt{1 - r_{12}^2}\, \sqrt{1 - r_{13}^2} \leqslant r_{23} \leqslant r_{12}\, r_{13} + \sqrt{1 - r_{12}^2}\, \sqrt{1 - r_{13}^2}\ . \quad \text{(i)}$$

Yule shows that the foregoing equation results from the fact that the partial correlation, $r_{23 \cdot 1}$, cannot exceed +1 and cannot be less than −1. This is true for any set of product-moment correlations (cf. 3, p. 330, and 8, pp. 238f.). The same conditions, as Burt has pointed out, can also be derived from the fact that the squared multiple correlation cannot exceed unity. Yule goes on to prove that analogous conditions obtain for the consistency of the correlations of between four or more variables : this can readily be deduced from the fact that $r_{12 \cdot 34}$, $r_{12 \cdot 345}$, etc., must lie between +1 and −1.

TABLE I. CORRELATIONS BETWEEN TRAITS

Trait		*V*10	*V*14	*V*8	*V*9	*S*3	*S*5	*S*6	*S*10	*C*8	*C*1	*C*6	*C*5
Greed for affection	*V*10	–	·81	·82	·79	—·39	—·33	—·38	—·31	—·45	—·35	—·40	—·36
Complacency	*V*14	·81	–	·75	·76	—·42	—·36	—·35	—·33	—·38	—·44	—·41	—·45
Sociophilia	*V*8	·82	·75	–	·79	—·30	—·37	—·41	—·38	—·52	—·44	—·35	—·37
Amiability	*V*9	·79	·76	·79	–	—·40	—·36	—·37	—·32	—·44	—·39	—·45	—·35
Energetic characteristic	*S*3	—·39	—·42	—·30	—·40	–	·74	·81	·75	—·49	—·71	—·60	—·59
Love domination..	*S*5	—·33	—·36	—·37	—·36	·74	–	·85	·78	—·67	—·58	—·67	—·56
Love risk and chance	*S*6	—·38	—·35	—·41	—·37	·81	·85	–	·82	—·68	—·69	—·59	—·59
Psychiatric callousness	*S*10	—·31	—·33	—·38	—·32	·75	·78	·82	–	—·67	—·63	—·60	—·69
Sociophobia	*C*8	—·45	—·38	—·52	—·44	—·49	—·67	—·68	—·67	–	·77	·84	·79
Restraint posture and movement ..	*C*1	—·35	—·44	—·44	—·39	—·71	—·58	—·69	—·63	·77	–	·79	·82
Emotional restraint	*C*6	—·40	—·41	—·35	—·45	—·60	—·67	—·59	—·60	·84	·79	–	·82
Apprehensiveness	*C*5	—·36	—·45	—·37	—·35	—·59	—·56	—·59	—·69	·79	·82	·82	–

So long as we are concerned solely with the consistency of Sheldon's figures, and not with their factorial analysis, we can disregard the *positive* correlations within each of the three main groups, *V*, *S*, and *C*, and scrutinize only the negative correlations between the different groups. In verbal form the problem might be stated thus : we can say that *two* things are the opposites each of the other ; when (if ever) can we say that *three* things are opposites ? The answer evidently is : only when the opposition is more or less incomplete. If opposition is expressed in quantitative form by a negative correlation, then (as the formula shows) three negative coefficients, all numerically larger than $-0{\cdot}50$, would be impossible. Accordingly, in Table I any negative values well over $-0{\cdot}60$ are likely to be highly suspect.

To obtain a decisive test, partial correlations involving three variables at a time were computed for each of the triads in Sheldon's correlation table that appeared suspicious. The well-known Pearson-Yule formula for partial correlation was employed, namely,

$$r_{12\cdot 3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}{}^2}\,\sqrt{1 - r_{23}{}^2}} \quad . \qquad . \qquad . \qquad .\text{(ii)}$$

III. RESULTS

Out of nineteen partial correlations computed in this manner, as many as five furnished negative coefficients numerically larger than -1. The detailed figures are set out below in Table II.

TABLE II. PARTIAL CORRELATIONS SHOWING IMPOSSIBLE VALUES

Correlated Traits	Trait Eliminated	Partial Correlations
*V*8, *S*5	*C*8	−1·13
*V*8, *S*6	*C*1	−1·09
*V*8, *S*6	*C*8	−1·22
*V*9, *S*3	*C*1	−1·04
*V*14, *S*3	*C*1	−1·16

It is evident that, if only one triad were found to give an impossible partial correlation, then at least one of the three first-order correlations would be in error. In the present case, the analysis suggests that, at the very least, three first-order correlations must involve errors of computation. With the data available it is impossible to determine with certainty which coefficients are in error and which may be safely accepted. The large negative values printed by Sheldon for the correlations between *S*3 and *C*1 (−0·71), *S*6 and *C*1 (−0·69), and *V*8 and *C*8 (−0·51) would account for all the anomalies. Or again, since the variables *V*8 and *C*1 occur three times, and one or other of the two appears in each of the five cases, mistakes over these two would suffice.

It will be noted that no question of linearity, sampling error, normality, homogeneous variance, etc., is involved. The matter is simply one of arithmetical accuracy.

The results emphasize the supreme importance of using methods of computation which include a complete system of checks (4, 7). A recent letter by Greenall in the *Journal* (2) makes it plain that it is not only tables of inter-correlations that can be inaccurate : errors can of course slip through even the most complex set of checks, but it is unlikely that large or numerous errors will result if one of the standard checking methods is adopted. Such methods should be treated as a matter of routine with all data intended for publication.

Finally, I wish to thank Mr. A. R. Jonckheere for assistance in the calculations made for this paper, and Professor Burt for guidance on several points, including the reference to Yule.

IV. SUMMARY

1. Difficulties have been experienced in factorizing Sheldon's table of correlations between temperamental traits. It is suggested that this is due to peculiarities, not in the underlying factors, but rather in the manner in which the correlations have been obtained: the pattern of negative coefficients is quite exceptional, and some of the coefficients appear to be arithmetically impossible.

2. This is confirmed by applying Yule's criterion,[1] namely, that, in an empirical table, no partial correlations can exceed unity. At least three of Sheldon's coefficients appear to be erroneously computed.

3. As a corollary it is suggested that all correlations should, as a matter of routine, be calculated by some method which permits of arithmetical checks.

REFERENCES

1. Adcock, C. J. (1948). 'A factorial examination of Sheldon's types.' *Journ. Pers.*, XVI, 312–19.
2. Greenall, P. D. (1949). 'Two criticisms.' *Brit. J. Psychol., Stat. Sect.*, II, 64–5.
3. Kendall, M. G. (1947). *The Advanced Theory of Statistics.* Vol. I. London: Griffin.
4. Kossack, Carl F. (1948). 'On the computation of zero-order correlation coefficients.' *Psychometrika*, XIII, 91–3.
5. McNemar, Quinn (1949). *Psychological Statistics.* New York: Wiley.
6. Sheldon, W. H. (1942). *The Varieties of Temperament.* New York: Harper.
7. Tucker, L. R. (1948). 'A note on the computation of a table of intercorrelations.' *Psychometrika*, XIII, 245–50.
8. Yule, Udny (1912). *An Introduction to the Theory of Statistics.* London: Griffin.

[1] In this article my chief purpose has been to demonstrate the importance of checking the arithmetical accuracy of the figures appearing in any table of inter-correlations and to show that those in Sheldon's table are mutually inconsistent. Professor Burt, however, points out that with other tables the criterion given above in equation (i) might be satisfied perfectly, and yet the coefficients remain inconsistent. "For example," he writes (in a personal communication), "a very slight change in Sheldon's table would suffice to reduce the partial correlations *of the first order* to an acceptable size; and yet, owing to the large negative correlations between Sheldon's *S*'s and *C*'s, it would still be impossible to analyse the table in terms of real factors." The more general criterion, that he has suggested elsewhere, consists in showing that *all the leading principal minors* are non-negative: when this has been proved, the matrix is positive definite (or semidefinite). These minors emerge automatically when the partial correlations are computed by what he calls the 'factorial method' (cf. *Factors of the Mind*, p. 307, and *Laboratory Notes on Elementary Statistics*, 1941, sect. 8): with this method therefore no very elaborate calculation would seem to be required.

I should like to take this opportunity of noting two errata in my previous article, this *Journal*, III, pp. 95 and 102. On p. 95 the plus sign under the radical should be minus; on p. 102 in the formula for c, read $\sqrt{N}$ for N.

BOOK REVIEWS

Human Ability. By C. Spearman and Ll. Wynn Jones. Macmillan & Company. Pp. vii + 198. 16s. net.

This slender volume embodies Professor Spearman's last work. It has been long awaited, and is addressed to all psychologists whether they happen to be interested in the statistical approach or not. Written in collaboration with Dr. Wynn Jones, it provides a comprehensive summary of Professor Spearman's final views on the factorial study of the mind—its methods, controversies, and results, and their psychological implications. The chapters suffer a little from the attempt to compress so much material into so small a space; and at times, in the effort to avoid excessive technicality, the writing drops into picturesque metaphor and loose colloquialisms, which often obscure the meaning instead of clarifying it. The manuscript was unfortunately left in an unfinished state; and, partly no doubt as a result, the text is somewhat inadequately annotated, and the references are not always complete or exact.

The problem with which the writers are chiefly concerned is the place to be allotted to 'group factors.' Those who are familiar with recent literature will be somewhat disappointed with the outcome of the discussion, for the views expressed are rather confused, and show little advance beyond the position adopted in *The Abilities of Man* some twenty years ago. There, it will be remembered, Professor Spearman, while still defending his celebrated theory of two factors, at the same time tried to leave a loophole for 'exceptional cases,' should they be finally confirmed.

In the present volume, we are assured, "the psychological picture still remains essentially the same." Nevertheless, in the summary to chapter II (headed "Theory of Two Factors") the writers tell us that "the statistical results of this method . . . consist of (1) a general factor; (2) an unlimited number of narrow specific factors; and (3) very few broad group factors." This formulation is almost identical, word for word, with the statement of the "three-factor theory," put forward by Burt over thirty years ago,[1] which Spearman at the time so strongly criticized. But how, the puzzled reader is tempted to enquire, can a two-factor method yield three kinds of factor? And, even if "very few" group factors are to be accepted, this nevertheless means that the two-factor theory, which postulates general and specific factors only, is no longer adequate.

Elsewhere, it is true, the authors try to set up a distinction between the kind of supplementary factors they are prepared to admit and the 'group factors' or 'special abilities' of earlier writers. Group factors, they say, must always depend on the tests put into the battery, and consequently cannot represent innate capacities. Their own factors they prefer to describe as 'broad *specific* factors' which, we are told, "must be sharply distinguished from what have been called *group* factors," although, a little inconsistently, they are termed 'group factors' in chapter II.

A new feature in their discussion is the development of a 'Method of Excess Correlation,' which is put forward as an alternative to existing group-factor methods. To illustrate the procedure, the writers print an artificial table of correlations arranged in such a fashion that all the highest correlations appear in the diagonal submatrices. This, of course, is the way in which an empirical correlation table is usually arranged before Burt's group-factor method is applied; and, as he has himself pointed out, the underlying principle goes back to Karl Pearson, who distinguished, in tables of correlations between physical measurements, two different levels of correlation, namely, the high positive correlations between 'like' traits and the comparatively low positive correlations between 'unlike' traits respectively. The general scheme had already been set out and criticized by Spearman and Hart in their well-known paper (*Brit. J. Psych.*, V, 1913, p. 57), where they claimed to show that it is *not* encountered in correlations between mental measurements. On this point, therefore, Spearman seems plainly to have changed his view. As regards the new method here proposed, an obvious defect arises from the fact that the formula necessarily assumes that all the cross-correlations, in which only the general factor is operative, are equal in numerical value.

---

[1] E.g., in *The Distribution and Relations of Educational Abilities* (1917) and the *Report on Psychological Tests of Educable Capacity* (1924).

Such an assumption is rarely fulfilled, even approximately, by the figures that are found in empirical tables.

Here and elsewhere neither the basic part played by Pearson in developing the product-moment formulæ, the principles of multiple correlation, and the method of principal axes, nor yet his many constructive criticisms, are ever explicitly mentioned. In an early chapter on "Some Criticisms," and later in the text, objections raised by Pearson and his followers are certainly cited, and refutations are offered; but Pearson himself is never referred to by name. This ban (as older psychologists in this country are well aware) dates from a heated controversy between the two which arose as far back as 1906. But by now it could surely be lifted. These anonymous references, to say the least, are highly misleading to the younger student.

To the statistical reader the most interesting section will doubtless be that which reviews and criticizes the methods of factor analysis developed by more recent workers. Spearman naturally takes first the 'bifactor method' put forward by Holzinger, since that claims to be based on the principles underlying his own two-factor theory: the method he accepts as valid in "certain conditions where the Two Factor analysis ceases to be applicable at all." He turns next to 'simple and weighted summation.' Here he makes quite clear his main objections to Burt's procedures. "Chronologically even earlier," he writes, "came another outstanding attempt at extending the scope of factorizing; this was contributed by Cyril Burt: whereas the older method dealt with every tetrad separately, the newer way took the sum of each whole column of correlations." All such attempts, however, he considers to be unsuccessful. His chief objection is to "the step whereby the vacant diagonal cells are filled out (by stopgaps) so as to make the whole table into a solid rectangular block." Thirdly, he briefly examines Thurstone's centroid method, and quotes his 'fundamental equation.' "This," he says, "is Burt's formula"; hence much the same criticisms are urged. And finally he glances at the method of 'principal components,' in which "the axes of the (correlation) ellipsoids correspond to the factors." This is attributed to Kelley and Hotelling, and no reference is made to the earlier outline of the procedure developed by Karl Pearson. It is summarily dismissed on the ground that, like the centroid method, it produces negative as well as positive saturations in the supplementary factors. Burt's bipolar factors Spearman appears to accept, at any rate in the temperamental field (p. 60). Up to now, he says, unlike the group and general factors, "they have attracted comparatively scant attention: but recently they have come more and more into their proper estate as research is turning to the other side of human nature—feeling and volition." In the cognitive field he rejects them, arguing that intellectual abilities cannot have a negative influence on actual performance. This criticism is plainly due to the fact that for Spearman a factor is always a causal factor, never a principle of classification. Consequently, he fails to understand that, in the view of the writers he is criticizing, when (say) a non-verbal test receives a negative saturation, what is negatived is the verbality, not its influence on actual performance. With Holzinger, he tells us, "the concept of 'bipolar factor' has been greatly extended"; and he thinks that "in this sphere of factor we must look to him for the fullest development."

What he calls 'indirect methods,' such as 'rotation' and the use of 'oblique axes,' are rejected because they too readily permit a subjective choice of whatever factors are required by the investigator's personal theories. His own method, he says, has been throughout that of partial correlation, with an *external* criterion rather than an internal. Taking external criteria (such as teachers' assessments for *g* or the Binet tests of intelligence) the general factor is partialled out by Yule's formula, and the residuum examined for supplementary factors. This, however, is scarcely factor analysis as now understood, since that proceeds primarily by seeking *internal* criteria derived from the table of inter-correlations themselves.

To those who knew Professor Spearman, it will be clear that he had lost none of his statistical ingenuity and little of his fighting spirit. To stress too strongly the shortcomings of the book would perhaps be unfair. As a pupil and devoted follower of Professor Spearman, Dr. Wynn Jones has evidently had a difficult task in preparing an incomplete manuscript for the press; and the labour which he has expended on the work shows how loyally he has carried it out. Everyone will feel deeply indebted to him for the way in which he has endeavoured to make Professor Spearman's final views available to the psychological reader.

E. K. C. B.

Psychological Statistics. By Quinn McNemar. New York: John Wiley & Sons, Inc. London: Chapman & Hall. 1949. Pp. viii + 364. $4.50.

Dr. McNemar is Professor of Psychology, Statistics, and Education at Stanford University; and his text-book is based on fifteen years' experience in teaching statistical procedures to students of education and psychology. It may be fairly said that it is probably the best and the most up to date of the many introductory works on the subject that have so far appeared on either side of the Atlantic. As Dr. McNemar has anticipated, the influence of his own teacher, Professor T. L. Kelley, is frequently discernible, and, as a result, that of Professor Karl Pearson, under whom Professor Kelley once studied. This perhaps makes its chapters specially suitable for psychological students in this country, since so many books and papers published by British psychologists have been, and still are, profoundly indebted to Karl Pearson's contributions. At the same time, the newer methods introduced by later writers, particularly by Dr. McNemar's other teacher, Professor Hotelling, are almost equally in evidence.

The manual is designed as a one-year course for senior and graduate students. Except for two occasions on which the calculus is used, only an elementary knowledge of mathematics is presupposed. The purpose is not so much to provide the student with a knowledge of techniques adequate for original research, but rather to enable him to understand the place and use of statistical concepts and procedures in current psychology.

Three chapters are devoted to the analysis of variance and covariance. Here the presentation seems decidedly superior to that adopted in most other elementary text-books; but its place in relation to the design of experiments might perhaps have been more clearly indicated. Further, no indication is given of the peculiar limitations of psychological data which so often prevent techniques like chi-square or analysis of variance from being taken over as they stand, and applied to problems of psychological research. Throughout the volume considerable space is given to demonstrative treatments of those topics in which the non-mathematical thinking is almost as difficult as the mathematical, for example, problems of sampling, and the notions of degrees of freedom and of interaction.

Nevertheless, the pure statistician, who has always been somewhat critical of the kind of statistical procedures favoured by psychologists, will probably have no difficulty in detecting here and there occasional inadequacies and even a little confusion in the discussion of certain logical problems. The usual distinction between large and small sample methods, which most statisticians would now consider somewhat arbitrary, is preserved. Possibly because they seem simple to the beginner, the methods of analysis appropriate to large samples are described first; but it is not made clear that these are really limiting cases of the methods appropriate to small samples, which are statistically more fundamental. Partly as a result of this approach, the questions of sampling, estimation, and the testing of statistical hypotheses, seem at times to get confused. Thus, in the introduction, the need for methods of statistical inference is said to arise " because of certain inadequacies of research data "; and the effect is to suggest that the major problem is one of simple estimation. The account of 'confidence' in connexion with statistical inference is restricted almost entirely to the estimation of the population mean; and there is no discussion of the more general problems of statistical inference as they relate to experimentation. The null hypothesis is introduced merely as " a better rule-of-thumb procedure "; and it is elsewhere stated, perhaps a little casually, that " chi-squared can likewise be used to test the null hypothesis." The discussion of experimental design makes only an incidental appearance.

As might be expected in a book intended for psychologists, considerable attention is devoted to multiple correlation and partial regression. Doolittle's method, by no means the easiest procedure for the beginner, is recommended for solving the requisite equations when four or more variables are involved. The treatment of rank-methods, which owe their introduction into practical statistical work so largely to psychologists, is brief and somewhat out of date: its limitations are stressed, but its advantages are unmentioned, and no hint is given of the recent work of Kendall on the subject. Since discriminant functions are closely related to methods of regression, and have begun to play an increasing part in psychological research, it is perhaps a little surprising that they too are wholly omitted. Factor analysis, we are told, has been deliberately excluded, presumably because even an introductory account was considered too difficult for the beginner. Thus the special statistical procedures which have been developed for the psychologist seem to be ignored;

and the psychological reader feels at times that the book is not so much an introductory review of *psychological* statistics as a means of introducing the student of psychology to techniques which have for the most part been developed in non-psychological fields.

The book makes a determined attempt to formulate the precise assumptions underlying the use of the various procedures as they are described, and to specify the conditions in which each method may be validly applied. This is well carried out in the discussion of correlation; but (in the reviewer's opinion) it is less successful in the explanations of the chi-square tests. The account given of the procedure to be followed in testing the difference between variances of independent random samples is not as complete or as exact as it should be: since tables of the variance ratio are, as a rule, appropriate only for the one-tail test required in analysis of variance, it is stated that for the present case the tabled probabilities should be doubled; but the author omits to point out that this procedure is only justifiable when the two samples are of equal size, nor indeed does he sufficiently bring out the importance of considering whether the assumption of homogeneity of variance is justified in particular cases. These, however, are blemishes which could perhaps be rectified in a later edition.

A. S.

Experimental Design in Psychological Research. By Allen L. Edwards. New York: Rinehart & Co. Inc. 1950. Pp. xvi + 446. $5.00.

Any text-book seeking to expound statistical procedures for psychological students has nowadays to keep a double aim in view. First, it should describe the more important techniques which are accepted by professional statisticians and which appear appropriate to the types of problem common in psychological research, even though such techniques have but rarely been adopted by psychologists. But secondly it must also explain those statistical methods which have in fact been in frequent use among psychologists themselves, and are therefore likely to be encountered in the books and periodicals consulted by the psychological student. If such methods have not yet been recognized by the professional statistician, or if they are now giving way to more rigorous procedures, these are not reasons for ignoring them altogether, but rather for explaining their imperfections.

This double task did not confront the earlier writers of such text-books, such as Thorndike, Kelley, or William Brown. The mathematical methods developed by Fechner and by the Galton-Pearson school were elaborated largely with psychological problems in mind. But most of the newer techniques in statistics have arisen out of work in very different fields. As mathematical devices they may be admirable in themselves; but too often they depend, explicitly or implicitly, theoretically or practically, on assumptions which may be valid enough in the field for which they were originally intended, but are not always appropriate to the data with which the psychologist has to deal.

The point is so frequently ignored that an illustration is perhaps desirable. In a recent study of tests applied to normal and neurotic groups (the latter reclassified into various subgroups) the investigator explains that he has applied the same kind of analysis as is to be found in Snedecor's chapter on the analysis of variance. He forgets that the distinction between the normal and the neurotic is by no means so clear-cut as between 'King' and 'Standard' varieties of apples, and that, had the measurements furnished by his psychological tests been repeated, the reliability would have been far lower than the reliability of the figures for the weight or the quantity of fruit. The critical psychologist is therefore tempted to ask whether the differences in the data may not invalidate the methods that have thus been borrowed, and whether they do not render these more refined working methods unnecessarily laborious. How far, for example, do the classifications and the experimental procedures of the psychologist yield anything that is precisely comparable to the 'blocks,' 'plots,' and 'treatments' which we encounter so frequently in *Statistical Methods for Research Workers?* I do not suggest that I myself share this view. But it is, I fancy, doubts like these that have made non-statistical psychologists in this country so sceptical of the statistical approach.

Unfortunately, admirable as they are, very few of the current text-books on statistical psychology really face these difficulties. In his latest volume, Professor Edwards has fulfilled the first of the two aims I have mentioned in a way which is beyond criticism. No doubt it is the more important of the two. But the second he seems almost wholly to overlook.

As the title suggests, his general approach follows that outlined by Fisher in *The Design of Experiments*. Hence a large part of the book deals with the analysis of variance and covariance and with the planning of factorial designs of varying degrees of complexity. A glance at the author index shows how deeply he is indebted to the work of British statisticians, like M. S. Bartlett, M. G. Kendall, F. Yates, Udny Yule, Karl Pearson, and Egon Pearson, as well as Fisher himself. And the number of actual researches on which he is able to draw in his search for concrete examples and illustrations shows how rapidly the work of such writers has begun to influence research-procedures in America. Hardly any, however, are taken from psychological experiments in this country. This is not because the work of British psychologists has been ignored, but because, until quite recently, it has so rarely adopted these newer techniques. Yet, if one turns back to some similar volume issued thirty years ago, one would find that nearly half the references relate to statistical work published by psychologists in this country. Today how many, among our senior colleagues in British psychological departments, could claim even a nodding acquaintance with " the newer developments in statistical analysis " due to British statisticians, and how many arrange for them to be taught to their research students ?

Nevertheless, I cannot help feeling that the attractive title chosen by Professor Edwards is a little too specific. Except for the fact that the examples are derived almost wholly from psychology, the title might almost as well have been phrased ' Experimental Design in *Statistical* Research.' Comparatively few of the methods described are characteristic of the typical or commoner kinds of experiment carried out in the psychological laboratory. The special statistical methods developed and used so persistently by psychologists themselves—for example, the constant process, the ranking method, the so-called discriminant methods, and factor analysis in its various branches—are not mentioned at all. Formerly statisticians were apt to retort, with Professor Filon, that " such devices are so amateurish that they should be discouraged rather than encouraged." Yet Kendall has shown that ranking procedures can be established on a sound basis ; Cramér's text-book includes references, not only to Pearson's method of principal components, but also to factor analysis by name ; while Fisher's own books include discussions of discriminant functions and the treatment of the sigmoid response curve by the probit method (i.e., what psychologists know as the constant process).

I think, too, it is a little unfortunate that the treatment might at first seem to imply that, if a psychological research is to be well planned, it must of necessity be statistical. The general principles of scientific methodology and of inductive argument are the same for all kinds of research ; and hence, except where a separate treatment is unavoidable on practical grounds, it is in my opinion a mistake, at any rate in text-books for the student of psychology, to divorce the designing of statistical investigations from the designing of scientific enquiries generally, whether statistical, or purely experimental, or merely observational.

However, it is scarcely fair to criticize an author for not discussing topics which he did not intend to take up. Professor Edwards may well reply that, at any rate for American students, these other matters are already sufficiently covered, and that his special purpose was to fill a gap that still needed filling. In his preface he explains that his primary object was " to present to the student in psychology, education, sociology, and the behavioural sciences some of the newer developments in statistical analysis, particularly with respect to small sample theory as they relate to the problems of research and experimentation in these fields." If this is taken to be his purpose, then it will be agreed that it has been excellently carried out. For students in this country who have already assimilated the older and more elementary techniques, perhaps by the aid of Professor Edwards' previous volume, there could be no better introduction to the " newer techniques." *C. B.*

The Backward Child. By Cyril Burt. London : University of London Press. Third Edition, 1950. Pp. xx + 704. 25s.

Sir Cyril Burt's book is concerned primarily with the causes and the treatment of backwardness among school children ; but it provides at the same time an admirable introduction to the study of individual differences among children of school age. Its chapters present in systematic form the results of a long series of investigations carried out by himself and by teachers working under his direction while he was Psychologist to the London County Council. To meet post-war changes, and to incorporate further knowledge accumulated during the last fifteen years, the whole volume has now been carefully revised, and the statistical sections considerably expanded.

In his earlier work on *The Young Delinquent* two new methodological features were introduced. The first consisted in an attempt to combine the case-study method with the statistical: the intention here was to use the quantitative approach to check the qualitative, and the qualitative to interpret the quantitative. The second consisted in the decision to carry out a parallel study on a control group, side by side with the study of the crucial group. The same two principles have been adopted in the investigation of educational backwardness.

In both cases the scheme underlying the examination of each child follows the same comprehensive lines, and is itself based on the results of prior statistical studies carried out by factorial methods. This is the portion of the work that has most frequently excited criticism, particularly from those who favour a different view of 'the factors of the mind.' Burt has always held that the apparent occurrence of both general and special disabilities among school children requires a fairly complex conception of mental processes—one which will include group factors as well as general factors. Not unnaturally those who set great store by "economy in psychological assumptions" were tempted to reply that one type of factor should surely be sufficient. To teachers outside training colleges and to practical psychologists concerned with the actual classification of subnormal children any such simplified explanations seemed scarcely workable. There was, however, an urgent need for a set of crucial enquiries, carried out by someone equally familiar with theoretical and with practical needs, along lines which would decide between the chief alternatives.

In this country, at the time when the volume was first published, Spearman's two-factor theory was still dominant. With a few dubious exceptions, so it was maintained, a single general ability would account for all individual differences; and the teacher's notion of special disabilities was dismissed as an out-of-date relic of the old faculty theory, which both general and statistical psychology had at length discarded. On the other hand, in America Thurstone and his followers had introduced the opposite picture: the hypothesis of a general factor was rejected, and the existence of group factors alone admitted. A somewhat similar standpoint had previously been suggested by Thomson.

The appointment of an official psychologist working in the London schools enabled researches to be planned and carried out on a scale that would have been quite impossible for the isolated investigator. The wide and varied data thus brought together in Burt's book appears to provide a conclusive answer to the issues raised: both the statistical analyses and the examination of numerous case-histories render the double hypothesis indispensable. We seem then bound to assume first an innate general ability, resembling what is commonly called intelligence, and secondly a number of innate abilities and disabilities in particular cognitive processes, such as memory, verbal capacity, and the like. The fact that, despite attacks from two opposite fronts, this view has now won increasing acceptance points plainly to the great methodological value of joining statistical with non-statistical modes of approach in all researches of this kind. Certainly, if one may judge from its fruitfulness in the present volume, Burt's hierarchical scheme of factors, including as it does both cognitive and orectic tendencies, with the broader or more general factors ramifying into narrower, lends itself admirably to the elucidation of the problems of individual psychology, at least so far as they arise in work among school children.

In discussing the place of quantitative methods in investigations of this kind, the author insists that statistical psychology is not to be treated as a distinct or self-contained branch: the use of quantitative procedures is "merely an extension or refinement of the logical techniques which are essential to all scientific research and argument." In the main, therefore, mathematical technicalities have been kept in the background; and in the text the critical ratios, tetrachoric correlations, and factor-saturations are merely cited incidentally in the course of the general argument to confirm or disprove the various alternative hypotheses.

During recent years, however, an increasing number of educationists, schoolmasters, and psychologists at child guidance centres have been undertaking statistical investigations on educational backwardness and kindred problems. Accordingly, in the present edition, the original appendices have been enlarged to supply a somewhat fuller account of the more useful computing devices and formulæ, and detailed references to recent publications have been added. But what should appeal to the statistical investigator most of all is the long and suggestive list of problems still awaiting systematic research.

P. B.

INDEX TO VOLUME III

INDEX OF AUTHORS

Authors	*Titles*	*Page*
Bartlett, M. S..	Tests of Significance in Factor Analysis	77–85
Burt, Cyril	Group Factor Analysis	40–75
Burt, Cyril	The Influence of Differential Weighting	105–125
Burt, Cyril	A Reply to Sir Godfrey Thomson's Note	127–128
Burt, Cyril	The Factorial Analysis of Qualitative Data	166–185
Henrysson, S.	The Significance of Factor Loadings	159–165
Lawley, D. N.	A Method of Standardizing Group-Tests	86–89
Lawley, D. N.	Factor Analysis by Maximum Likelihood : A Correction	76
Lubin, A.	Linear and Non-Linear Discriminating Functions	90–104
Lubin, A.	A Note on Sheldon's Table of Correlations between Temperamental Traits	186–189
McLeish, J.	The Validation of Seashore's Measures of Musical Talent by Factorial Methods	129–140
Raath, M. J., and Reyburn, H. A.	Primary Factors of Personality	150–158
Reid, D. D.	Night Visual Capacity and its Relation to Survival in Operational Flying	141–149
Reyburn, H. A., and Raath, M. J.	Primary Factors of Personality	150–158
Richardson, L. F.	A Method for Computing Principal Axes	16–20
Sen, Amya	A Statistical Study of the Rorschach Test	21–39
Thomson, Godfrey	Note on Sir Cyril Burt's Paper on Differential Weighting	126
Vernon, P. E.	An Application of Factorial Analysis to the Study of Test Items	1–15

INDEX OF SUBJECTS

Titles	*Authors*	*Page*
Differential Weighting, The Influence of	Burt, Cyril	105–125
Differential Weighting, Note on Sir Cyril Burt's Paper on	Thomson, Godfrey	126
Differential Weighting, A Reply to Sir Godfrey Thomson's Note	Burt, Cyril	127–128
Discriminating Functions, Linear and Non-Linear	Lubin, A.	90–104
Factor Analysis by Maximum Likelihood : A Correction	Lawley, D. N.	76
Factor Analysis, Tests of Significance in	Bartlett, M. S.	77–85
Factor Loadings, The Significance of	Henrysson, S.	159–165
Group Factor Analysis	Burt, Cyril	40–75
Group Tests, A Method of Standardizing	Lawley, D. N.	86–89
Night Visual Capacity and its Relation to Survival in Operational Flying	Reid, D. D.	141–149
Personality, Primary Factors of	Reyburn, H. A., and Raath, M. J.	150–158
Principal Axes, A Method for Computing	Richardson, L. F.	16–20
Qualitative Data, The Factorial Analysis of	Burt, Cyril	166–185
Rorschach Test, A Statistical Study of the	Sen, Amya	21–39
Seashore's Measures of Musical Talent, The Validation of, by Factorial Methods	McLeish, J.	129–140
Sheldon's Table of Correlations between Temperamental Traits, A Note on	Lubin, A.	186–189
Test Items, An Application of Factorial Analysis to the Study of	Vernon, P. E.	1–15

BOOKS REVIEWED

Author	*Title*	*Page*
Burt, C.	The Backward Child	194–195
Edwards, Allen L.	Experimental Design in Psychological Research	193–194
McNemar, Q.	Psychological Statistics	192–193
Spearman, C., and Jones, Ll. Wynn	Human Ability	190–191

RESEARCH BUREAU

EAU OF EDNL. & PSYL. RESE